BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification: Not classified	A2	(11) International Publication Number: **WO 90/07862**
		(43) International Publication Date: 26 July 1990 (26.07.90)

(21) International Application Number: PCT/US90/00024

(22) International Filing Date: 5 January 1990 (05.01.90)

(30) Priority data:
293,980 6 January 1989 (06.01.89) US

(71) Applicant: PROTEIN ENGINEERING CORPORATION [US/US]; 765 Concord Avenue, Cambridge, MA 02138 (US).

(72) Inventors: LADNER, Robert, Charles ; 3827 Green Valley Road, Ijamsville, MD 21754 (US). GUTERMAN, Sonia, Kosow ; 20 Oakley Road, Belmont, MA 02178 (US). KENT, Rachel, Baribault ; 51 Houghton Road, Wilmington, MA 01887 (US). LEY, Arthur, Charles ; 122 Adena Road, Newton, MA 02165 (US).

(74) Agent: COOPER, Iver, P.; Browdy and Neimark, 419 7th Street, N.W., ,300, Washington, DC 20004 (US).

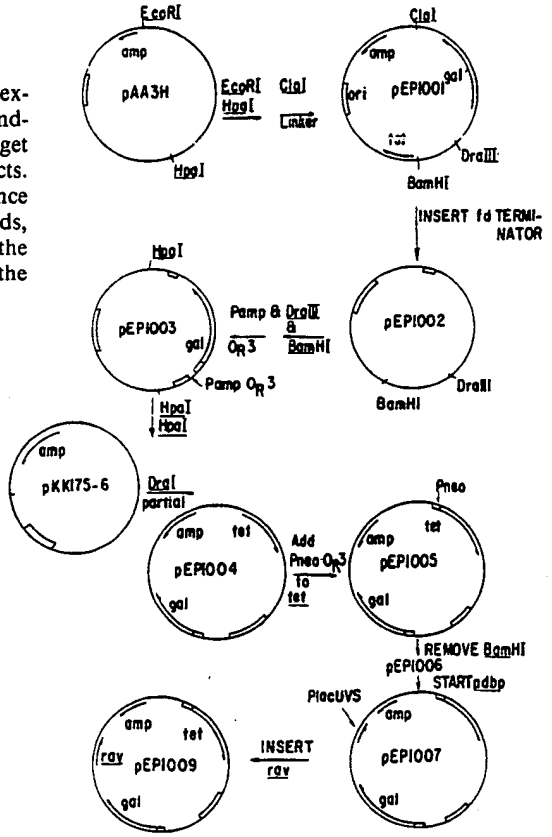
(81) Designated States: AT, AT (European patent), AU, BB, BE (European patent), BF (OAPI patent), BG, BJ (OAPI patent), BR, CA, CF (OAPI patent), CG (OAPI patent), CH, CH (European patent), CM (OAPI patent), DE, DE (European patent), DK, DK (European patent), ES (European patent), FI, FR (European patent), GA (OAPI patent), GB, GB (European patent), HU, IT (European patent), JP, KP, KR, LK, LU (European patent), MC, MG, ML (OAPI patent), MR (OAPI patent), MW, NL, NL (European patent), NO, RO, SD, SE, SE (European patent), SN (OAPI patent), SU, TD (OAPI patent), TG (OAPI patent).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: GENERATION AND SELECTION OF NOVEL DNA-BINDING PROTEINS AND POLYPEPTIDES

(57) Abstract

Novel DNA-binding proteins, especially repressors of gene expression, are obtained by variegation of genes encoding known binding proteins and selection for proteins binding the desired target DNA sequence. A novel selection vector is used to reduce artifacts. Heterooligomeric proteins which bind to a target DNA sequence which need not be palindromic are obtained by a variety of methods, e.g., variegation to obtain proteins binding symmetrized forms of the half-targets and heterodimerization to obtain a protein binding the entire asymmetric target.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MR	Mauritania
BE	Belgium	GA	Gabon	MW	Malawi
BF	Burkina Fasso	GB	United Kingdom	NL	Netherlands
BG	Bulgaria	HU	Hungary	NO	Norway
BJ	Benin	IT	Italy	RO	Romania
BR	Brazil	JP	Japan	SD	Sudan
CA	Canada	KP	Democratic People's Republic of Korea	SE	Sweden
CF	Central African Republic			SN	Senegal
CG	Congo	KR	Republic of Korea	SU	Soviet Union
CH	Switzerland	LI	Liechtenstein	TD	Chad
CM	Cameroon	LK	Sri Lanka	TG	Togo
DE	Germany, Federal Republic of	LU	Luxembourg	US	United States of America
DK	Denmark	MC	Monaco		

GENERATION AND SELECTION OF NOVEL DNA-BINDING PROTEINS AND POLYPEPTIDES

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to development of novel DNA-binding proteins and polypeptides by an iterative process of mutation, expression, selection, and amplification. The ability to create novel DNA-binding proteins will have far-reaching applications, including, but not limited to, use in: a) treating viral diseases, b) treating genetic diseases, c) preparation of novel biochemical reagents, and d) biotechnology to regulate gene expression in cell cultures. Several workers have shown that repressors derived from bacteria function when expressed in eukaryotic cells (BREN84, FIGG88, BROW87, HUMC87, HUMC88), but none have shown how to generate proteins that bind sequence-specifically to a predetermined DNA sequence. For reviews of transcriptional control in eukaryotic cells, see STRU87, JONE87, and MANI87. The present application deals only with sequence-specific DNA-binding proteins, abbreviated DBP.

Proteins, particularly repressors, having affinity for specific sites on DNA modulate transcription of genes. The best known are a group of proteins primarily studied in prokaryotes that contain the structural motif alpha-helix-turn-alpha-helix (H-T-H) (SAUE82, PABO84). These proteins bind as dimers or tetramers to DNA at specific operator sequences that have approximately palindromic sequences. Contacts made by two adjacent alpha helices of each monomer in and around two sites in the major groove of B-form DNA are a major feature in the DNA-protein interface. This group of proteins includes phage repressor and

Cro proteins, bacterial metabolic repressors such as GalR, LacI, LexA, and TrpR, bacterial activator protein CAP and activator/repressor AraC, bacterial transposon and plasmid TetR proteins (PAB084), the yeast mating type regulators MATa1 and MATalpha2 (MILL85) and eukaryotic homeo box proteins (EVAN88).

Interactions between dimeric repressors and approximately palindromic operators have usually been discussed in the literature with attention focused on one half of the operator with the tacit or explicit assumption that identical interactions occur in each half of the complex. Departures from palindromic symmetry allow proteins to distinguish among multiple related operators (SADL83, SIMO84). One must view the DNA-protein interface as a whole. The emphasis in the literature on dyad symmetry is a barrier to determining the requirements for general novel recognition of DNA by proteins.

The equilibrium geometry and flexibility of DNA are determined by the sequence; see inter alia HOGA87, GART88, and ULAN87. The interactions of ionic, polar, and hydrophobic groups on the DNA with solvent molecules and ions make detailed predictions of DNA conformation and binding properties very difficult; cf. OHLE85, ULAN87, and OTWI88.

Matthews (MATT88), commenting on the current collection of protein-DNA structures, concludes that: a) different H-T-H DBPs use their recognition helices differently, b) there is no simple code that relates particular base pairs to particular amino acids at specific locations in the DBP, and c) "full appreciation of the complexity and individuality of each complex will be discouraging to anyone hoping to find simple answers to the recognition problem." Schleif (SCHL88) has characterized the study of DNA-binding proteins as a field still in its infancy and emphasizes the difficulties of designing proteins that bind predetermined sequences.

Prokaryotic repressors exist that are unrelated to H-T-H binding proteins. Some of these bind to approximate palindromic sequences (e.g. Salmonella typhimurium phage P22 Mnt protein (VERS87a) and E. coli TyrR repressor protein (DEFE86)). Others bind to operator sequences that are partially symmetric (S. typhimurium phage P22 Arc protein, VERS87b; E. coli Fur protein, DELO87; plasmid R6K pi protein, FILU85) or non-symmetric (phage Mu repressor, KRAU86).

Genetics has enabled extensive analysis of prokaryotic DNA-binding proteins and their specific nucleic acid recognition sequences. It is not yet possible, however, to design a protein to bind strongly and specifically to an arbitrary DNA sequence. As taught by the present invention it is, nonetheless, possible to postulate a family of potential DBP mutants and identify one having the desired specificity by other means.

Genetic studies of the DNA-binding proteins show that mutations in protein sequence that result in decrease of protein function fall into two overlapping classes: 1) those that destabilize the global protein structure or folding and 2) those that specifically alter the binding properties. The first class illuminates the general problem of protein folding and stability, while the second defines the interactions involved in the formation and stabilization of the protein-DNA complex. Mutations in the operator yield additional information.

Positions 84 to 91 in helix 5 of λ repressor have been subjected to extensive amino acid substitutions (REID88). Two or three positions were varied simultaneously through all twenty amino acids and those combinations giving normal function were selected. The authors neither discuss optimization of the number or positions of residues to vary to obtain any particular functionality, nor did they

attempt to obtain proteins having alternate dimerization or recognition functions.

Pakula et al. (PAKU86) have randomly mutagenized λ Cro. They sought and found non-functional mutants but did not seek or find proteins having novel DNA-binding properties, nor did they suggest how to select such proteins.

Sequence-independent DNA-protein interactions are thought to occur via electrostatic interactions between the backbone of the DNA and charged or polar groups of the protein (ANDE87, LEWI83, and TAKE85). Sequence-specific interactions involve H-bonding, nonpolar, or van der Waals contacts between exposed side groups or groups of the polypeptide main chain and base pair edges exposed in the major and minor grooves of the DNA.

Mutations that alter residues involved in specific binding interactions with DNA have been identified in prokaryotic DBPs, including λ, 434, and P22 repressor and Cro proteins, P22 Arc and Mnt, and E. coli trp and lac repressors and CAP. These mutations occur in residues that are exposed to solvent in the free protein but buried in the protein-DNA complex.

A few cases have been reported (BASS88, YOUD83, VERS85a, CARU87, WHAR85b, WHAR87, EBRI84, and SPIR88) in which a change in one or a few residues in a DNA-binding protein not only abolishes binding by the protein to the wild-type operator but also confers strong binding to a different operator. In all the cited publications, alteration of binding specificity has been accomplished by using symmetrically-located pairs of alterations in the operator sites. Single, asymmetric changes or multiple changes asymmetrically located in either the binding protein or its operator were not considered. In "helix swap" experiments (WHAR84, WHAR85b, WHAR85a, SPIR88, BUSH88, PABO84), multiple mutations are introduced into the

DNA-binding recognition helix of H-T-H proteins with the goal of changing the operator specificity of one known DBP to that of a different known DBP.

An extension of the "helix swap" experiments uses a mixture of 434 repressor and 434R[alpha3(P22R)] (HOLL88). This mixture recognizes and binds in vitro with high affinity to a 16 bp chimeric operator comprising a 434 half-site and a P22 half-site, indicating that active heterodimers are formed. The authors did not extend the results to intracellular repression, nor did they perform mutagenesis of the repressors and selection of cells to create novel recognition patterns.

Two approaches have been developed to create novel proteins through reverse genetics. In one approach, dubbed "protein surgery" (DILL87), a substitution is introduced at a single protein residue. This approach has been used to determine the effects on structure and function of specific substitutions in trypsin (CRAI85, RAOS87, BASH87).

The other approach has been to generate a variety of mutants at many loci within the cloned gene, the "gene-directed random mutagenesis" method. The specific location and nature of the change or changes are determined post hoc by DNA sequencing. If loss of a wild-type function confers a cellular phenotype, one screens colonies for mutations; see, cf PAKU86. This approach is limited by the number of colonies that can be examined. An additional important limitation is that many desirable protein alterations require multiple amino acid substitutions and thus are not accessible through single base changes or even through all possible amino acid substitutions at any one residue.

The objective in both these approaches has been, however, to analyze the effects of a variety of point mutations, so that rules governing such substitutions could

be developed (ULME83). Progress has been hampered by the efforts involved in using either method (ROBE86).

Oliphant et al. (OLIP86) and Oliphant and Struhl (OLIP87) have demonstrated ligation and cloning of highly degenerate oligonucleotides and have applied saturation mutagenesis to the study of promoter sequence and function. They have suggested that similar methods could be used to study genetic expression of protein coding regions of genes, but they do not say how one should: a) choose protein residues to vary, or b) select or screen mutants with desirable properties.

Ward et al. (WARD86) have engineered heterodimers from homodimers of tyrosyl-tRNA synthetase. Methods of converting homodimeric DBPs into heterodimeric DBPs are disclosed in the present invention. Methods of deriving single-polypeptide pseudo-dimeric DBPs from homodimeric DBPs are disclosed in the examples of the present invention.

Benson et al. (BENS86) have developed a scheme to detect genes for sequence-specific DNA-binding proteins. They do not consider non-symmetric target DNA sequences nor do they suggest mutagenesis to generate novel DNA-binding properties. Their method is presented as a method to detect genes for naturally occurring DNA-binding proteins. Because the selective system is lytic growth of phage, low levels of repression can not be detected. Selective chemicals, as disclosed in the present application, on the other hand, can be finely modulated so that low level repression is detectable.

Elledge and Davis (ELLE89a) and Elledge et al. (ELLE89b) have used an occluded aadA gene in a selection for cells expressing eukaryotic DBPs. The supposed recognition sequence of the sought DBP is incorporated into the strong promoter that occludes aadA on a low-copy number

plasmid. Their system is presented as a tool for cloning pre-existing DBPs and there is no mention of variegation of the gene that encodes the potential DBP. Furthermore, there is no discussion of the symmetry of the target sequence or of the symmetry of the DBP.

Ladner and Bird, WO88/06601 suggest strategies for the preparation of asymmetric repressors. In one embodiment, a gene is constructed that encodes, as a single polypeptide chain, the two DNA-binding domains of a naturally-occurring dimeric repressor, joined by a polypeptide linker that holds the two binding domains in the necessary spatial relationship for binding to an operator. While they prefer to design the linker based on protein structural data (cf. Ladner, U.S. Patent 4,704,692) they state that uncertainties in the design of the linker may be resolved by generating a family of synthetic genes, differing in the linker-encoding subsequence, and selecting in vivo for a gene encoding the desired pseudo-dimer. Ladner and Bird do not consider the background of false positives that would arise if the two-domain polypeptides dimerize to form pseudo-tetramers.

The binding of lambdoid repressors, Cro and CI repressor, is taken, in WO88/06601, as canonical even though other DBPs were known having operators of different lengths. WO88/06601 maintains that the 17 bp lambdoid operators can be divided into three regions: a) a left arm of five bases, b) a central region of seven bases, and c) a right arm of five bases. Several other DBPs are known for which this division is inappropriate. Further, WO88/06601 states that the sequence and composition of the central region, in which edges of bases are not contacted by the DBP, are immaterial. There is direct evidence for 434 repressor (KOUD87, KOUD88) that the sequence and composition of the central region strongly influences binding of 434 repressor.

Once a pseudo-dimer is obtained, they then obtain an asymmetric pseudo-dimer by the following technique. First, the user of WO88/06601 is directed to construct a family of hybrid operators in which the sequence of the left and right arms are specified; no specification is given for the central seven bases. In each member of the family, the left arm contains the same sequence as the wild-type operator left arm while the right arm 5-mer is systematically varied through all 1024 possibilities. Similarly, in the gene encoding the pseudodimer, the codons for one recognition helix have the wild-type sequence while the codons coding for the other recognition helix are highly varied. The variegated pseudodimer genes are expressed in bacterial cells, wherein the hybrid operators are positioned to repress a single highly deleterious gene. Thus, it is supposed that one can identify a recognition helix for each possible 5-mer right arm of the operator by in vivo selection; the correspondences between 5-mer right arms and sequences of recognition helices are compiled into a dictionary. The consequences of mutations or deletions in the deleterious genes are not considered. WO88/06601 suggests that successful constructions may be very rare, e.g. one in 10^6, but ignore other genetic events of similar or greater frequency.

To obtain a repressor for an arbitrary 17-mer operator, the user of WO88/06601:

a) finds the 5-mer sequence of the left arm in the dictionary and uses the corresponding recognition helix sequence in the first DNA-binding domain of the pseudodimer,

b) ignores the sequence and composition of the next seven bases, and

c) finds the 5-mer sequence of the right arm in the dictionary and uses the corresponding recognition

helix sequence in the second DNA-binding domain of the pseudodimer.

WO88/06601 also envisions means for producing a heterodimeric repressor. A plasmid is provided that carries genes encoding two different repressors. A population of such plasmids is generated in which some codons are varied in each gene. WO88/06601 instructs the user to introduce very high levels of variegation without regard to the number of independent transformants that can be produced. WO88/06601 also instructs the user to introduce variegation at widely separated sites in the gene, though there is no teaching concerning ways to simultaneously introduce high levels of variegation at widely separated sites in the gene or concerning maintenance of diversity without selective pressure, as would be needed if the variegation were introduced stepwise. WO88/06601 teaches that codons thought to be involved in the protein-protein interface should be preferentially mutated to generate heterodimers. Cells transformed with this population of plasmids will produce both the desired heterodimer and the two "wild-type" homodimers. WO88/06601 advises that one select for production of the heterodimer by providing a highly deleterious gene controlled by a hybrid operator, and beneficial genes controlled by the wild-type operators. The fastest growing cells, it is taught, will be those that produce a great deal of the heterodimer (which blocks expression of the deleterious gene) and little of the homodimers (so that the beneficial genes are more fully expressed). There is no consideration of mutations or deletions in the deleterious gene or in the wild-type operators; such mutations will produce a background of fast-growing cells that do not contain the desired heterodimers.

SUMMARY OF THE INVENTION

This invention relates to the development of novel proteins or polypeptides that preferentially bind to a specific subsequence of double-stranded DNA (the "target") which need not be symmetric, using a novel scheme for in vivo selection of mutant proteins exhibiting the desired binding specificities.

The novel binding proteins or polypeptides may be obtained by mutating a gene encoding on expression: 1) a known DNA-binding protein within the subsequence encoding a known DNA-binding domain, 2) a protein that, while not possessing a known DNA-binding activity, possesses a secondary or higher order structure that lends itself to binding activity (clefts, grooves, helices, etc.), 3) a known DNA-binding protein but not in the subsequence known to cause the binding, or 4) a polypeptide having no known 3D structure of its own.

This application uses the term "variegated DNA" to refer to a population of molecules that have the same base sequence through most of their length, but that vary at a number of defined loci. Using standard genetic engineering techniques, variegated DNA can be introduced into a plasmid so that it constitutes part of a gene (OLIP86, OLIP87, CHEN88, AUSU87, REID88). When plasmids containing variegated DNA are used to transform bacteria, each cell makes a version of the original protein. Each colony of bacteria produces a different version from most other colonies. If the variegations of the DNA are concentrated at loci that code on expression for residues known to be on the surface of the protein or in loops, a population of genes will be generated that code on expression for a population of proteins, many members of which will fold into roughly the same 3D structure as the parental protein. Most often we generate mutations that are concentrated within codons for residues thought to make contact with the DNA. Secondari-

ly, we introduce mutations into codons specifying residues that are not directly involved in DNA contact but that affect the position or dynamics of residues that do contact the DNA.

In general, a variegated population of DNA molecules, each of which encodes one of a large (e.g. 10^7) number of distinct potential target-binding proteins, is used to transform a cell culture. The cells of this cell culture are engineered with binding marker genetic elements so that, under selective conditions, the cell thrives only if the expressed potential target-binding protein in fact binds to the target subsequence preventing transcription of these binding marker genetic elements. (Typically, binding of a successful target-binding protein to the target subsequence blocks expression of a gene product that is deleterious under selective conditions. Alternatively, binding of a successful target-binding protein can inactivate a strong promoter that otherwise occludes transcription of a beneficial gene.) The mutant cells are directed to express the potential target-binding proteins and the selective conditions are applied. Cells expressing proteins binding successfully to the target are thus identified by in vivo selection. If the binding characteristics are not fully satisfactory, the amino acid sequences of the best binding proteins are determined (usually by sequencing the corresponding genes), a new population of DNA molecules is synthesized that encode variegated forms of the best binding proteins of the last cull, mutant cells are prepared, the new population of potential DNA-binding proteins is expressed, and the best proteins are once again identified by the superior growth of the corresponding transformants under selective conditions. The process is repeated until a protein or polypeptide with the desired binding characteristics is obtained. Its corresponding gene may then be moved to a suitable expression system.

In the simplest form of this invention, the mutant cells are provided with a selectable genetic element, the transcription of which is deleterious to the survival or growth of the cell. The selectable genetic element either is a promoter or is operably linked to a promoter regulating the expression of the gene. The promoter, or other non-coding region of the genetic element (for example, an intron), has been modified to include the desired target subsequence in a position where it will not interfere with transcription of the selectable gene unless a protein binds to that target subsequence. Each mutant cell is also provided with a gene encoding on expression a potential DNA-binding protein, operably linked to a promoter that is preferably regulated by a chemical inducer. When this gene is expressed, the potential DNA-binding protein has the opportunity to bind to the target and thereby protect the cell from the selective conditions under which the product of the binding marker gene would otherwise harm the cell.

In addition to the desired outcome of these in vivo selections, there exist a number of possible genetic events that allow the cells to escape the selection, producing artifacts and inefficiency by allowing the growth of colonies that do not express the desired sequence-specific DNA-binding proteins. Examples of mechanisms, other than the desired outcome, that lead to cell survival under the selective conditions include: a) a point mutation or a deletion in the selectable gene eliminates expression or function of the selectable gene product; b) a host chromosomal mutation compensates for or suppresses function of the selectable gene product; c) the introduced potential DNA-binding protein binds to a DNA subsequence other than the chosen target subsequence and blocks expression of the selectable gene; d) the introduced potential DNA-binding protein binds to and inactivates the gene product of the selective gene; and e) a DNA-binding protein endogenous to the host mutates so that it binds to the selectable gene and blocks expression of the selectable gene.

This invention relates, in particular, to the design of a vector that confers upon the host cells the desired conditional sensitivity to the selection conditions in such a manner as to greatly reduce the likelihood of false positives and artifactual colonies.

First, at least two selectable genes that are functionally unrelated are used to reduce the risk that a single point mutation in the vector (or in the host chromosome) will destroy the sensitivity of the cell to the selective conditions, since it will eliminate only one of the two (or more) deleterious phenotypes. Similarly, a single introduced gene for a potential DNA-binding protein that binds to and inactivates the gene product of one selectable gene will not bind and inactivate the gene product of the other selectable gene. The likelihood that point mutations will occur in both selectable genes or that two host chromosomal mutations will spontaneously arise that suppress the effects of two genes is the product of each single individual probabilities of the necessary event, and thus is extremely low.

The DNA sequences of the two or more selectable genes preferably should not have long segments of identity: a) to avoid isolation of a DBP that binds these identical regions instead of the intended target sequence, and b) to reduce the likelihood of genetic recombination. The degeneracy of the genetic code allows us to avoid exact identity of more than a few, e.g. 10, bases.

Second, the selectable genes are placed on the vector in alternation with genetic elements that are essential to plasmid maintenance. Thus, a single deletion event, even of thousands of bases, cannot eliminate both selectable genes without also eliminating vital genetic elements. Alternatively, the selectable genes are placed in the

bacterial chromosome. Spontaneous deletions from the chromosome are rare.

Third, different promoters are associated with each of the selectable genes. This ensures that the selection does not isolate cells harboring genes encoding on expression novel DNA-binding proteins that bind specifically to subsequences that are part of the promoter but not the chosen target subsequence. Each cell expresses only one or a few introduced potential DNA-binding proteins (multiple potential DNA-binding proteins could arise if one cell is transformed by two or more variegated plasmids). The probability that two such proteins will occur in one cell and that one will bind to the promoter of the first selectable gene and that the second will bind to the different promoter of the second selectable gene is very small.

Fourth, the selectable binding marker genes may be placed on a vector different from the vector that carries the potential dbp gene. DNA manipulations that introduce variegation into the potential dbp gene can cause mutations in the vector remote from the site of the intended mutations. Thus, we may place the selectable binding marker genes in the bacterial chromosome or on a separate plasmid that is compatrible with the dbp vector.

Finally, the same promoter is used to initiate transcription of two genes: a) one of the deleterious selectable binding marker genes, and b) a beneficial or essential gene also borne on the plasmid and used to select for uptake and maintenance of the plasmid (e.g. an antibiotic resistance gene, such as bla). In the case of the beneficial or essential gene, however, there is no instance of the predetermined target DNA subsequence associated with the promoter. Thus, if a DNA-binding protein binds to a subsequence of the promoter other than the predetermined target DNA subsequence, it will frustrate

expression of the beneficial or essential one. If desired, more than one such beneficial or essential gene may be provided. In that event, copies of promoter A may be operably linked to both deleterious gene A' (with an instance of the target) and beneficial gene A" (without an instance of the target), while copies of promoter B are operably linked to both deleterious gene B' (with target) and beneficial gene B" (without target).

The selection system described above is a powerful tool that eliminates most of the artifacts associated with selections based on cloning vectors that use a single selectable gene or that have all selectable genes in a contiguous region of the plasmid. While this invention embraces using the aforementioned elements of a selection system singly or in partial combination, most preferably all are employed.

In one embodiment, the invention relates to a cell culture comprising a plurality of cells, each cell bearing:

i) a gene coding on expression for a potential DNA-binding protein or polypeptide, where such protein or polypeptide is not the same for all such cells, but rather varies at a limited number of amino acid positions; and

ii) at least two independent operons, each comprising at least one binding marker gene coding on expression for a product conditionally deleterious to the survival or reproduction of such cells, the promoter of each said binding marker gene containing a predetermined target DNA subsequence so positioned that, if said target DNA subsequence is bound by a DNA-binding protein or polypeptide, said conditionally deleterious product is not expressed in functional form.

Most known DNA-binding proteins bind to palindromic or nearly palindromic operators. It is desirable to be able to obtain a protein or polypeptide that binds to a target DNA subsequence having no particular sequence symmetry. In another embodiment of the present invention, such a binding protein is obtained by creating a hybrid of two dimeric DNA-binding proteins, one of which (DBP_L) recognizes a symmetrized form of the left subsequence of the target subsequence, and the other of which (DBP_R) recognizes a symmetrized form of the right subsequence of the target subsequence.

Cells producing equimolar mixtures of DBP_L and DBP_R contain approximately 1 part $(DBP_L)_2$, 2 parts DBP_L:DBP_R, and 1 part $(DBP_R)_2$. The DBP_L:DBP_R heterodimers, which bind to the non-symmetric target subsequence, may be isolated from a cell lysate by affinity chromatography using the target sequence as the ligand. If desired, the heterodimers may be stabilized by chemically crosslinking the two binding domains.

It is also possible to modify both DBP_L and DBP_R, by a process of variegation and selection, so that they have (without disturbing their affinity for the predetermined DNA target subsequence) complementary but not dyad-symmetric protein-protein binding surfaces. When such polypeptides are mixed, in vivo or in vitro, the primary species will be DBP_L:DBP_R heterodimers. Alternatively, reversing the steps, a dimeric binding protein may be modified so that its two binding domains have complementary but not dyad-symmetric protein-protein binding surfaces, and then the DNA-contacting surfaces are modified to bind to the right and left halves of the target DNA subsequence. In either case, the resulting cooperative domains can be crosslinked for increased stability.

When a binding protein is engineered so that its two binding domains have complementary, but not dyad-symmetric

protein-protein binding surfaces, then in the preferred embodiment one of the steps will be a "reverse selection", i.e. a s lection for a protein that does not bind to the symmetrized half-target sequence. To facilitate such reverse selection, it is desirable that the binding marker genes be capable of "two-way" selection (VINO87). For a two-way selectable gene there exist both a first selection condition in which the gene products are deleterious (preferably lethal) to the cell and a second selection condition in which the gene product is beneficial (preferably essential) to the cell. The first selection condition is used for forward selection in which we select for cells expressing proteins that bind to the target so that gene expression is repressed. The second selection condition is used for reverse selection in which we select for cells that do not express a protein that binds to the target, thereby allowing expression of the gene product.

Abolition of function is much easier than engineering of novel function. Reverse selection can isolate cells that: a) express no DBP, b) express unstable proteins descendant from a parental DBP, c) express a protein descendant from a parental DBP having very nearly the same 3D structure as the parental DBP, but lacking the functionality of the parent. We are interested in this third class. It is difficult, however, to distinguish among these classes genetically. Therefore, when using reverse selection, we carefully choose sites to mutate the protein (so as to minimize the chances of destroying tertiary structure) and we introduce a lower level of variegation than in forward selection. We must verify biochemically that a stable, folded protein is produced by the isolated cells.

Another concept of the present invention is the use of a polypeptide, rather than a protein, to preferentially bind DNA. This polypeptide, instead of binding the DNA molecule as a preformed molecule having shape complementary

to DNA, will wind about the DNA molecule in the major or minor groove. Such a polypeptide has the advantages that: a) it is smaller than a protein having equivalent recognizing ability and may be easier to introduce into cells, and b) it may serve as a model for creation of other compounds that bind DNA sequence-specifically.

In a preferred embodiment, transcription of the DNA that codes on expression for potential-DNA-binding proteins or polypeptides is regulated by addition of chemical inducer to the cell culture, such as isopropylthiogalactoside (IPTG). Other regulatable promoters having different inducers or other means of regulation are also appropriate.

The invention encompasses the design and synthesis of variegated DNA encoding on expression a collection of closely related potential DNA-binding proteins or polypeptides characterized by constant and variable regions, said proteins or polypeptides being designed with a view toward obtaining a protein or polypeptide that binds a predetermined target DNA subsequence.

For the purposes of this invention, the term "potential DNA-binding polypeptide" refers to a polypeptide encoded by one species of DNA molecule in a population of variegated DNA wherein the region of variation appears in one or more subsequences encoding one or more segments of the polypeptide having the potential of serving as a DNA-binding domain for the target DNA sequence or having the potential to alter the position or dynamics of protein residues that contact the DNA. A "potential DNA-binding protein" (potential-DBP) may comprise one or more potential DNA-binding polypeptides. Potential-DBPs comprising two or more polypeptide chains may be homologous aggregates (e.g. A_2) or heterologous aggregates (e.g. AB).

From time to time, it may be helpful to speak of the "parental sequence" of the variegated DNA. When the novel DNA-binding domain sought is a homolog of a known DNA-binding domain, the parental sequence is the sequence that encodes the known DNA-binding domain. The variegated DNA is identical with this parental sequence at most loci, but will diverge from it at chosen loci. When a potential DNA-binding domain is designed from first principles, the parental sequence is a sequence that encodes the amino acid sequence that has been predicted to form the desired DNA-binding domain, and the variegated DNA is a population of "daughter DNAs" that are related to that parent by a high degree of sequence similarity.

The fundamental principle of the invention is one of forced evolution. The efficiency of the forced evolution is greatly enhanced by careful choice of which residues are to be varied. The 3D structure of the potential DNA-binding domain and the 3D structure of the target DNA sequence are key determinants in this choice. First a set of residues that can either simultaneously contact the target DNA sequence or that can affect the orientation or flexibility of residues that can touch the target is identified. Then all or some of the codons encoding these residues are varied simultaneously to produce a variegated population of DNA. The variegated population of DNA is introduced into cells so that a variegated population of cells producing various potential-DBPs is obtained.

The highly variegated population of cells containing genes encoding potential-DBPs is selected for cells containing genes that express proteins that bind to the target DNA sequence ("successful DNA-binding proteins"). After one or more rounds of such selection, one or more of the chosen genes are examined and sequenced. If desired, new loci of variation are chosen. The selected daughter genes of one generation then become the parental sequences for the next generation of variegated DNA (vgDNA).

DNA-binding proteins (DBPs) that bind specifically to viral DNA so that transcription is blocked will be useful in treating viral diseases, either by introducing DBPs into cells or by introducing the gene coding on expression for the DBP into cells and causing the gene to be expressed. In order to develop such DBPs, we need use only the nucleotide sequence of the viral genes to be repressed. Once a DBP is developed, it is tested against virus in vivo. Use of several independently-acting DBPs that all bind to one gene allow us to: a) repress the gene despite possible variation in the sequence, and b) to focus repression on the target gene while distributing side effects over the entire genome of the host cell. Animals, plants, fungi, and microbes can be genetically made intracellularly immune to viruses by introducing, into the germ line, genes that code on expression for DBPs that bind DNA sequences found in viruses that infect the animal (including human), plant, fungus, or microbe to be protected.

Sequence-specific DBPs may also be used to treat autoimmune and genetic disease either by repressing noxious genes or by causing expression of beneficial genes.

Some naturally-occurring DBPs bind sequence-specifically to DNA only in the presence of absence of specific effector molecules. For example, Lac repressor does not bind the lac operator in the presence of lactose or isopropylthiogalactoside (IPTG); Trp repressor binds DNA only in the presence of tryptophan or certain analogues of tryptophan. The method of the present invention can be used to select mutants of such DBPs that a) recognize a different cognate DNA sequence, or b) recognize a different effector molecule. These alterations would be useful because: a) known inducible or de-repressible DBPs allows us to use the novel DBP without affecting existing metabo-

lic pathways. Having novel effectors allows us to induce or de-repress the regulated gene without altering the state of genes that are controlled by the natural effectors. In addition, temperature-sensitive DBPs could be made which would allow us to control gene expression in the same way that λ cI857 and P_R and P_L are used.

Conferring novel DNA-recognition properties on proteins will allow development of novel restriction enzymes that recognize more base pairs and therefore cut DNA less frequently. For example, the methods of the present invention will be useful in developing a derivative of EcoRI (recognition GAATTC) that recognizes and cleaves a longer recognition site, such as TGAATTCA. Proteins that recognize specific DNA sequences may also be used to block the action of known restriction enzymes at some subset of the recognition sites of the known enzyme, thereby conferring greater specificity on that enzyme. Other DNA-binding enzymes may also be obtained by the methods described herein.

The methods of the present invention are primarily designed to select from a highly variegated population those cells that contain genes that code on expression for proteins that bind sequence-specifically to predetermined DNA sequences. The genetic constructions employed can also be used as an assay for putative DBPs that are obtained in other ways.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 Schematic of protein bound to DNA.
Figure 2 Schematic of evolution of a binding protein.
Figure 3 Plasmid pKK175-6.
Figure 4 Plasmid pAA3H.
Figure 5 Summary of construction of pEP1009.
Figure 6 Plasmid pEP1001.
Figure 7 Plasmid pEP1002.

Figure 8 Plasmid pEP1003.
Figure 9 Plasmid pEP1004.
Figure 10 Plasmid pEP1005.
Figure 11 Plasmid pEP1007.
Figure 12 Plasmid pEP1009.

DETAILED DESCRIPTION OF THE INVENTION AND ITS PREFERRED EMBODIMENTS

Abbreviations :

The following abbreviations will be used throughout the present invention:

Abbreviation	Meaning
DBP	DNA-binding protein
idbp	A gene encoding the initial DBP
pdbp	A gene encoding a potential-DBP
vgDNA	variegated DNA
dsDNA	double-stranded DNA
ssDNA	single-stranded DNA
Tc^R, Tc^S	Tetracycline resistance or sensitivity
Gal^R, Gal^S	Galactose resistance or sensitivity
Gal^+, Gal^-	Ability or inability to utilize galactose
Fus^R, Fus^S	Fusaric acid resistance or sensitivity
Km^R, Km^S	Kanamycin resistance or sensitivity
Ap^R, Ap^S	Ampicillin resistance or sensitivity

--- *** ---

Terminology

A domain of a protein that is required for the protein to specifically bind a chosen DNA target subsequence, is referred to herein as a "DNA-binding domain". A protein may comprise one or more domains, each composed of one or more polypeptide chains. A protein that binds a DNA sequence specifically is denoted as a "DNA-binding protein". In one embodiment of the present invention, a preliminary operation is performed to obtain a stable protein, denoted as an "initial DBP", that binds one specific DNA sequence. The present invention is concerned with the expression of numerous, diverse, variant "potential-DBPs", all related to a "parental potential-DBP" such as a known DNA-binding protein, and with selection and amplification of the genes encoding the most successful mutant potential-DBPs. An initial DBP is chosen as parental potential-DBP for the first round of variegation. Selection isolates one or more "successful DBPs". A successful DBP from one round of variegation and selection is chosen to be the parental DBP to the next round. The invention is not, however, limited to proteins with a single DNA binding domain since the method may be applied to any or all of the DNA binding domains of the protein, sequentially or simultaneously.

Amino acids are indicated by the single-letter code, AUSU87, Appendix A.

Symbols that represent ambiguous DNA are: T, C, A, G for themselves; M for A or C; R for A or G; W for A or T; S for C or G; Y for T or C; K for G or T; V for A, C, or G; H for A, C, or T; D for A, G, or T; B for C, G, or T; N for any base.

Conventionally, DNA sequences are written from 5' to 3', left-to-right.

```
anti-sense DNA: 5' ATG CTT TTC ... 3'
     sense DNA: 3' TAC GAA AAG ... 5'

          mRNA: 5' AUG CUU UUC ... 3'

       protein:     M - L - F -
```

We will use the convention that the "sense" strand is the strand used as template for mRNA synthesis.

In the present invention, the words "grow", "growth", "culture", and "amplification" mean increase in number, not increase in size of individual cells. In the present invention, the words "select" and "selection" are used in the genetic sense; i.e. a biological process whereby a phenotypic characteristic is used to enrich a population for those organisms displaying the desired phenotype.

One selection is called a "selection step"; one pass of variegation followed by as many selection steps as are needed to isolate a successful DBP, is called a "variegation step". The amino acid sequence of one successful DBP from one round becomes the parental potential-DBP to the next variegation step. We perform variegation steps iteratively until the desired affinity and specificity of DNA-binding between a successful DBP and chosen target DNA sequence are achieved.

In a "forward selection" step, we select for the binding of the PDBP to a target DNA sequence; in a "reverse selection" step, for failure to bind. The target DNA sequence may be the final target sequence of interest, or the immediate target may be a related sequence of DNA (e.g., a "left symmetrized target" or "right symmetrized target"). There is an important distinction between

screening and selection. Screening merely reveals which cells express or contain the desired gene. Selection allows desired cells to grow under conditions in which there is little or no growth of undesired cells (and preferably eliminates undesired cells).

The term "operon" is used to mean a collection of one or more genes that are transcribed together. We will use operon to refer also to one or more genes that are transcribed together in eukaryotic cells independent of post-transcriptional processing.

The term "binding marker gene" is used to mean those genes engineered to detect sequence-specific DNA binding, as by association of a target DNA with a structural gene and expression control sequences. A single operon may include more than one binding marker gene (e.g., galT,K). A "control marker gene" is one whose expression is not affected by the specific binding of a protein to the target DNA sequence. The "control promoter" is the promoter operably linked to the control marker gene.

Palindrome, palindromic, and palindromically are used to refer to DNA sequences that are the same when read along either strand, e.g.

Palindromic DNA

Rotational axis

↓

5' C T A G C C T A G G C T A G 3'
3' G A T C G G A T C C G A T C 5' .

The arrow indicates the center of the palindrome; if the sequence is rotated 180° about the central dot, it appears unchanged. In the present application, "Palindromic" does not apply to sequences that have mirror symmetry within one strand, such as

```
                Mirror Plane
                     |
5' C T A G C C T|T C C G A T C 3'
3' G A T C G G A|A G G C T A G 5'  .
                     |
```

DNA sequences can be partially palindromic about some point (that can be either between two base pairs or at one base pair) in which case some bases appear unchanged by a 180° rotation while other bases are changed.

A special case of partially palindromic sequence is a "gapped palindrome" in which palindromically related bases are separated by one or more bases that lack such symmetry:

Gapped Palindrome

```
   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
5' C T A G C T T T C C G G C T A G 3'
3' G A T C G A A A G G C C G A T C 5'
```

has CTAGC (bases 1-5) palindromically related to GCTAG (bases 12-16) while the sequence TTTCCG (bases 6-11) in the center has no symmetry.

For the purposes of this invention, a "non-deleterious cloning site" is a region on a plasmid or phage that can be cut with one restriction enzyme or with a combination of restriction enzymes so that a large linear molecule comprising all essential elements can be recovered.

Overview: Standard Methods

Bacterial strains are cultured by standard methods (DAVI80, MILL72, AUSU87). Constructions of vectors are by standard methods (MANI82, ZOLL84, AUSU87). All genetic constructions are confirmed, first by analysis with restriction enzymes, and then by sequencing. Sequencing is by the Sanger dideoxy method or by Maxam Gilbert chemical

method. Constructions that confer a phenotype are tested for display of the desired phenotype. These necessary controls are not described repeatedly.

Overview: The Selection System

The present invention separates mutated genes that specify novel proteins with desirable sequence-specific DNA-binding properties from closely related genes that specify proteins with no or undesirable DNA-binding properties, by: 1) arranging that the product of each mutated gene be expressed in the cytoplasm of a cell carrying a chosen DNA target subsequence, and 2) using genetic selections incorporating this chosen DNA target subsequence to enrich the population of cells for those cells containing genes specifying proteins with improved binding to the chosen target DNA sequence.

A selectably deleterious gene is positioned relative to, usually downstream from, the target sequence so that the gene is not expressed if a successful DNA-binding protein specific to this target is expressed in the cell and binds the target sequence. Alternatively, a selectable beneficial gene may be arranged so that its transcription is occluded by a strong promoter (ADHY82, ELLE89a, ELLE8-9b). The target sequence is placed in or near the occluding promoter so that successful binding by a protein will repress the occluding promoter and allow transcription of the beneficial gene. Elledge and coworkers disclose that such systems work best in the bacterial chromosome or on low-copy-number plasmids. The cell will survive exposure to the selective conditions transcription of the selectably deleterious genetic element is blocked.

The preferred cell line or strain is easily cultured, has a short doubling time, has a large collection of well characterized selectable genes, includes variants that are deficient in genetic recombination, and has a well devel-

oped transformation system that can easily produce at least 10^7 independent transformants/ug of DNA. Bacterial cells are preferred over yeasts, fungi, plant, or animal cells because they are superior on every count. Among bacteria, E. coli is the premier candidate because of the wealth of knowledge of genetics and cellular processes. Other bacterial strains, such as S. typhimurium, Pseudomonas aeruginosa, Klebsiella aerogenes, Bacillus subtilis, or Streptomyces coelicolor could be used. DBPs that bind to host regulatory sequences, such as promoters, will be toxic. Thus, development of a DBP that specifically binds to E. coli promoters is preferably done in a cell line or strain, such as S. coelicolor, having significantly different promoter sequences.

In the most preferred embodiments, all novel DBPs are developed in E. coli recA$^-$ strains. The recA$^-$ genotype is preferred over other rec$^-$ mutations because recA$^-$ mutation reduces the frequency of recombination more than other known rec$^-$ mutations and the recA$^-$ mutation has fewer undesirable side effects. We choose a host strain that methylates or does not methylate the target sequence in the desired way. For example a Dcm$^-$ strain is appropriate if the target sequence contains CCwGG and we want a DBP that binds the unmethylated form.

As vectors, phage, such as M13, have the advantage of a high infectivity rate. Organisms or phage having a phase in their life cycle in which the genome is single-stranded DNA have a higher mutation rate than organisms or phage that have no phase in which the genome is single-stranded DNA. Plasmids are, however, preferred because genes on plasmids are much more easily constructed and altered than are genes in the bacterial chromosome and are more stable than genes borne on phage, such as M13. M13 derived vectors are nearly as preferred as plasmids.

In some embodiments, the cloning vector will carry: a) the selectable genes for successful DBP isolation, b) the pdbp gene, c) a plasmid origin of replication, and d) an antibiotic resistance gene not present in the recipient cell to allow selection for uptake of plasmid. Preferably the operative vector is of minimum size.

Alternatively, the selectable binding marker genetic elements are placed on a vector different from but compatible with the vector that carries the pdbp gene. This arrangement has the advantages that engineering the pdbp gene is easier on a smaller plasmid and manipulation of pdbp can not introduce mutations into the selectable binding marker genes.

Standard selections for plasmid uptake and maintenance in E. coli include use of antibiotics (e.g. ampicillin (Ap)) as shown in Table 2. Selection of cells with antibiotics is preferred to nutritional selections, e.g. $TrpA^+$, for several reasons. Nutritional selection may be overcome by large volumes of cells or growth medium; host chromosomal auxotrophy is rarely total; crossfeeding of the non-growing cells by prototrophic recipients obscures the outlines of the colonies; and late mutations to prototrophy may arise on the plate due to spontaneous mutation of nongrowing cells. Nonetheless, nutritional selection may be employed.

Similarly, plasmids for use in B. subtilis are engineered for selection of uptake and maintenance using antibiotics. Plasmids used in streptomycete species bear genes for resistance to antibiotics such as thiostrepton, neomycin, and methylenomycin, in preference to auxotrophic markers or sporulation and pigment screens such as spo in bacilli and mel in streptomycetes.

Recombinant DNA manipulations in yeasts have been achieved using complementation of auxotrophic markers, some

of which are shown in Table 3. High backgrounds are surmounted by use of two unrelated binding marker genes carried on the same vector, e.g., Leu2$^+$ and Ura3$^+$. Selection for G418 resistance conferred by the bacterial aphII gene expressed in yeast offers the advantages of reduced background and a wider range of appropriate recipient strains. The current upper range of efficiency of DNA uptake into yeast cells indicates that this organism is not now preferred for the process described in this patent, although results could be achieved by large scale practice.

The selection systems must be so structured that other mechanisms for loss of gene expression are much less likely than the desired result, repression at the target DNA subsequence. Other mechanisms that could yield the desired phenotype include: point mutations that inactivate the deleterious gene or genes, deletion of the deleterious gene or genes, host mutations that suppress the deleterious genes, and repression at a site other than the target DNA sequence.

A wide range of selectable phenotypes for E. coli and S. typhimurium have been described (VINO87). Two broad classes of selections are useful in this invention, nutritional and chemical. Such selections are inherently conditional in that they employ addition of a growth-inhibitory chemical to the selective medium, or manipulation of the nutrient components of the selective medium. Further conditionality of the preferred method is imposed by transcriptional regulation (e.g. by IPTG in combination with the lacUV5 promoter and the LacI^q repressor) of the variegated pdbp gene. In those members of the population that express DBPs that bind to the target, IPTG indirectly controls the selectable genes; in these cells, increased IPTG leads to reduced expression of the selectable genes. Therefore the phenotypes for selection are distinguished only in the presence of an inducing chemical, and potential

deleterious effects of these phenotypes are avoided during storage and routine handling of the strains.

Selection of mutant strains capable of producing proteins that can bind to the target DNA subsequence is enabled by engineering conditional lethal genes or growth-inhibiting genes located downstream from the promoter that contains the target DNA subsequence. In the preferred embodiment, at least two independent conditional lethal or inhibitory selections are performed simultaneously. It is possible to use a single selection to achieve the same purpose, but this is not preferred. Two selections are strongly preferred since a simple mutation in the selected gene, occurring at a frequency of 10^{-6} to 10^{-8}/cell, would occur in two selected genes simultaneously at the product of the individual frequencies, 10^{-12} to 10^{-16}. Thus use of two selections substantially reduces the probability of isolation of artifactual revertant or suppressor strains.

Selectable genes for which both forward and reverse selections exist are preferred because, by changing host or media, we can use these genes to select for binding by a DBP to a target DNA sequence such that expression of one of these genes is repressed, or we can select phenotypes characteristic of cells in which there is no binding of the DBP. For example, expression of the tet gene is essential in the presence of tetracycline. On the other hand, expression of the tet gene is lethal in the presence of fusaric acid. Expression of the galT and galK genes in a $GalE^-$ host in the presence of galactose is lethal (NIKA61). Expression of galT and galK in a host that is $GalE^+$ and either $GalT^-$ or $GalK^-$ renders the cells Gal^+ and allows them to grow on galactose as sole carbon source.

The term "source of a selective agent" includes the selective agent itself and any media components which cause the cell to manufacture the selective agent.

The Detailed Examples describe selection of strains with successful DBP binding to novel target subsequences due to turn off of two genes, each of which, if expressed, confers sensitivity to a toxic substance. It is also possible to use selection of strains in which successful DBP binding to novel target operators turns off repressors of genes encoding required gene products. For example, using the binding marker gene P22 arc, we place an Arc operator site so that binding of Arc represses expression of a beneficial or conditionally essential gene, such as amp. Another alternative is selection of expression of required gene products due to successful binding of DBP proteins derived from positive effectors as the DBP, e.g. CAP from E. coli, the repressor from phage λ, or the Cro67 (BUSH88) mutant of λ Cro. Another alternative is to place the target sequence in or near a strong promoter that occludes transcription of a conditionally essential gene (ELLE89a,b).

The selections described in the Detailed Examples employ commercially available cloned genes on plasmids in strains that can be obtained from the ATCC (Rockville, MD). Alternatively, the genes can be produced synthetically from published sequences or isolated from a suitable genomic or cDNA library.

Numerous types of selections are possible for selection of DBP expression in E. coli. The toxic and inhibitory agents listed in Table 4 are used with appropriately engineered host strains and vectors to select loss of gene function listed above. Repression of transcription of these genes allows growth in the presence of the agents. Other outcomes such as deletions or point mutations in these genes may also be selected with these agents, hence two functionally unrelated selections are used in combination. These agents share the property that cell metabolism is stopped, and unlike the nutritional selections, the inhibitory agents are not overcome by components of the

growth medium or turnover of macromolecules in the cells. Selections using antibiotics, metabolite analogs, or inhibitors are preferred. Another class of selections includes those for repression of phage or colicin receptors, or for repression of phage promoters. These agents kill by single-hit kinetics, and in the case of phage, are self-replicating, making the multiplicity of agent to putative repressed cell much more difficult to control and so are not preferred (BENS86).

Any selection system relevant to the cell line or strain may be substituted for those in the examples given here, with appropriate changes in the engineering of the cloning vectors. One example is the dominant pheS$^+$ gene carried on plasmid pHE3 (ATCC #37,161) in a pheS12 background. Turn-off of pheS$^+$ is selected with p-fluorophenylalanine (Sigma Corp., St. Louis, MO).

We could choose the Streptomyces coelicolor cloned glucose kinase gene for selection of the DBP$^+$ phenotype, using the metabolite analog deoxyglucose.

Each batch of antibiotic is checked for MIC (minimum inhibitory concentration) under the condition of use. Increased concentration of antibiotic may be used to increase the stringency of the selectio, in most cases.

The user varies the medium formulation (pH, cation concentrations, buffering agent, etc.) for a particular selection if the results are not optimal with the strain at hand. For example, Maloy and Nunn (MALO81) describe a medium yielding improved selection of Fus^R E. coli colonies from a Tc^R background, compared to the medium employed by Bochner (BOCH80) for this purpose using S. typhimurium.

Stringency of selection can be modulated by controlling copy number of plasmids bearing the selectable genes;

increasing copy number of selectable genes increases the stringency of the selection.

During the initial phases of the progressive development of DBP molecules, it is desirable to produce a high intracellular concentration of DBP. The stringency of the selection is increased in subsequent phases of successful DBP development by allowing fewer molecules of DBP per cell. Thus it is preferred to regulate transcription of pdbp by an inducible or derepressible promoter, such as PlacUV5.

High total cell input often decreases stringency of selections, by providing metabolites that are specifically omitted, by mass action with respect to an inhibitory agent, or by generating a large number of artificial satellite colonies that follow the appearance of genetically resistant colonies. The number of cells that are successfully transformed is a function of efficiency of ligation and transformation processes, both of which are optimized in the embodiment of this invention. Procedures for maximal transformation and ligation efficiency are from Hanahan (HANA85) and Legerski and Robberson (LEGE85) respectively. Increasing stringency is imposed under the conditions of high efficiency of these processes by inoculation of plates with small volumes or dilutions of cell samples. Pilot experiments are performed to determine optimum dilution and volume.

In Detailed Example 1, the transformation event is followed by dilution and growth of cells in permissive medium following transformation. Exogenous inducer of DBP expression is included at this step, and a set of selections are then imposed in liquid medium. Surviving cells are concentrated by centrifugation, and selected for these and additional traits using solid medium in Petri plates. This protocol offers th advantage that fewer identical siblings are obtained and a larger population is easily

screened. In Detailed Example 1, repression of the Gal^S phenotype is selected by exposing transformants to galactose in liquid medium, which produces visible lysis of galactose sensitive cells. The second selection employed in Detailed Example 1 is for the Fus^R phenotype due to repression of Tc^R, which requires limitation of total inoculum size to 10^6 cells/plate. Similar protocol variations are introduced to combine selections for transformation and successful DBP function.

Tests of selective agents to determine the conditions that kill or inhibit sensitive cells are performed with pure cultures of sensitive cells. These include strains carrying the selective marker genes having the recognition sequence of the IDBP as target, with and without idbp, and with and without the inducer of idbp expression.

Cultures of sensitive cells are applied to selective media as inocula appropriate to the selection (usually 10^6 to 10^8 per plate). Sufficient numbers of replicates (10^7 to 10^9 total sensitive cells for each medium) are tested by each selection. The rate at which the cultures produce revertants and phenotypic suppressors (considered together as revertants) is determined. A rate greater than 10^{-6} per cell indicates that stringency must be increased. If reversion rates are below this level, as we have shown for the selections described in Example 1, mixing experiments are performed to determine the sensitivity of recovery of a small fraction of resistant cells from a vast excess of sensitive cells.

Normally, the deleterious gene product of a binding marker gene is a protein. It may also be an RNA, e.g., an mRNA which is antisense to the mRNA of an essential gene and therefore blocks translation of the latter mRNA into protein. Another alternative is that transcription of the binding marker gene may be deleterious b cause this transcription occludes transcription of an adjacent

beneficial gene. Selectively deleterious genes suitable for use in the present invention include those shown in Table 4.

The two selectably deleterious genes are preferably not functionally related. For example, the chosen genes should not code for proteins localized to or affecting the same macromolecular assembly in the cell or which alter the same or intersecting anabolic or catabolic pathways. Thus, use of two inhibitors that select for mutations affecting RNA synthesis, aromatic amino acid synthesis, or each of histidine and purine synthesis are not preferred. Similarly, two inhibitors that are transported into the cell by shared membrane components are thus functionally related, and are not preferred. In this manner the user reduces the frequency of isolation of single host mutations that yield the apparent desired phenotype, because of suppression of the shared functionality, interacting component, or precursor relationship. Host mutations of this type are conveniently distinguished by a screen of the selectable phenotypes in the absence of the inducer of the DBP, e.g. IPTG.

Examples of pairs of deleterious genes which are recommended for use in the present invention are given in Table 5A. In each case, one of the paired genes codes for a product that acts intracellularly while the other codes for a product that acts either in transport into or out of the cell or acts in an unrelated biological pathway. Table 5B gives some pairs that are not recommended. These pairs have not been shown to malfunction, but they are not recommended, given the large number of choices that are clearly functionally unrelated.

A preferred novel feature is the use of a copy of the promoter of one of these beneficial or conditionally essential genes, operably linked to the target DNA subsequence, to direct transcription of the selectably deleter-

ious or conditionally lethal binding marker genes of the plasmid. If the potential-DBP should repress the selectable gene by binding to this promoter, it would also repress this beneficial activity.

In order to assure that selection for DBP binding is specific to the target and not the promoter, we, preferably, place one of the two selectable binding marker genes under the same transcription initiation signal as the gene we use for selection of vector maintenance. In Detailed Example 1, transcription of the galT and galK genes is initiated by the P_{amp} promoter, as is the amp gene.

It is possible that the potential-DBP will bind specifically to the boundary between the target DNA sequence and the promoter, or within the structural gene. In the preferred embodiment, we discriminate against this mechanism by choosing a different promoter, operably linked to another copy of the same target DNA sequence, for the second selectable gene. Preferably, the two promoters that initiate transcription of the selectable genes should be strong enough to give a sensitive selection, but not too strong to be repressed by binding of a novel DBP. Some well studied promoters and their scores by the Mulligan algorithm (MULL84) are shown in Table 6. Promoters that score between 50% and 70% are good candidates for use in binding marker genes. Preferably, the two promoters have significant sequence differences, particularly in the region of the junction to the target DNA sequence. Specifically, the region between the -10 region and the target sequence, which comprises five to seven bases, should have no more than two identical bases in the two promoters. Although the -10 regions of promoters show high homology, promoters are known (e.g. P_{amp} having GACAAT and P_{neo} having TAAGGT) that have as few as two out of six bases identical in this region, and such difference is preferred.

The target DNA sequence for the potential DNA-binding protein must be associated with the two deleterious binding marker genes and their promoters so that expression of the binding marker genes is blocked if a novel protein in fact binds to the target sequence. The target DNA sequence could appear upstream of the gene, downstream of the gene, or, in certain hosts, in a noncoding region (viz. an "intron") within the gene. Preferably, it is placed upstream of the coding region of the gene, that is, in or near the RNA polymerase binding site for the gene, i.e. the promoter. If the binding marker gene is an occluding promoter, the target is, preferably, placed downstream of the promoter. Placement of the target DNA sequence relative to the promoter is influenced by two main considerations: a) protein binding should have a strong effect on transcription so that the selection is sensitive, b) the activity of the promoter in the absence of a binding protein should be relatively unaffected by the presence of the test DNA sequence compared to any other target subsequence.

In the present invention, we will deal primarily with DNA target subsequences of 10 to 25 bases. It has been noted that the highly conserved -35 region and the highly conserved -10 region are separated by between 15 and 21 base pairs with a mode of 17 base pairs (HAWL83, MULL84). Some of the bases between -35 and -10 are statistically non-random; thus placement of target DNA sequences longer than 10 bases between the -10 and -35 regions would likely affect the promoter activity independent of binding by potential-DBPs. Because quantitative relationships between promoter sequence and promoter strength are not well understood; it is preferable, at present, to use known promoters and to position the target at the edge of the RNA polymerase binding site.

Protein binding to DNA has maximum effect on transcripti n if the binding site is in or just down-stream

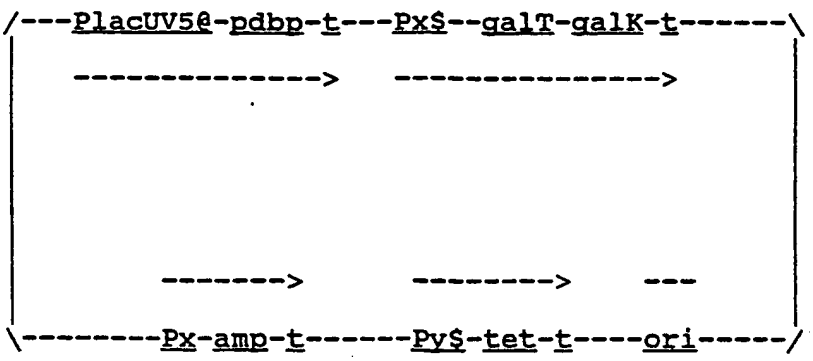
from the promoter of a gene. Hoopes and McClure (HOOP87) have reviewed the regulation of transcription initiation and report that the LexA binding site can produce effective repression in a variety of locations in the promoter region. In a preferred embodiment, we place the target DNA sequences that begin with A or G so that the first 5' base of the target sequence is the +1 base of the mRNA, as the LexA binding site is located in the uvrD gene (HOOP87, p1235). If the target sequence begins with C or T, we preferably place the target so that the first 5' base of the target is the +2 base of the mRNA and we place an A or G at the +1 position. An alternative is to place the target DNA sequences upstream of the -35 region as the LexA binding site is located in the ssb gene (HOOP87, p1235).

It may be useful in early stages of the development of a DBP to have more than one copy of the target DNA sequence positioned so that binding of a DBP reduces transcription of the selectable gene. Multiple copies of the target DNA sequence enhances the sensitivity of phenotypic characteristics to binding of DBPs to the target DNA sequence. Multiple copies of the target DNA sequence are, preferably, placed in tandem downstream of the promoter. Alternatively, one could place one copy upstream of the promoter and one or more copies downstream.

We arrange the genes on the plasmid or plasmids in such a way that no single deletion event eliminates both deleterious genes without also eliminating a gene essential either to plasmid replication or cell survival. Thus, resistant colonies are unlikely to arise through deletions because two independent deletion events are required. Similarly, simultaneous occurrence of one point mutation and one deletion is as unlikely as two point mutations or two deletions.

A typical arrangement of genes on the operative

cloning vector, similar to that used in Detailed Example 1, is:



Px represents the promoter that initiates transcription of the amp gene. A second copy of Px initiates transcription of galT,K. Py is a promoter driving tet, t is a transcriptional terminator (different terminators may be used for different genes), and $ is the target subsequence. PlacUV5 is the lacUV5 promoter, @ represents the lacO operator, and pdbp is a variegated gene encoding potential DBPs. Placement of the pdbp relative to other genes is not important because mutations or deletions in pdbp cannot cause false positive colony isolates. Indeed, it is not necessary that the pdbp gene be on the selection vector at all. The purpose of the selection vector is to ensure that the host cell survives only if the one of the PDBPs binds to the target sequence (forward selection) or fails to so bind (reverse selection). The pdbp gene may be introduced into the host cell by another vector.

Two-way selections are available for both tet and galT,K (vide supra). The orientation of each gene in the selection vector is unimportant because strong terminators (e.g. rrnBt1, rrnBt2, phage fd terminator) are preferably placed at the ends of each transcription unit. That galT,K and tet are separated by essential genes, however, is of fundamental imp rtance. The sequence ori is essential for plasmid replication, and the amp gene, the transcription of

which is initiated by Px, is essential in the presence of Ap. Successful repression of galT,K and tet is selected with galactose and fusaric acid. No single deletion event can remove both the latter genes and allow plasmid maintenance or cell survival under selection. In addition, binding by a novel DBP to the Px promoter would render the cell Ap sensitive. These arrangements make appearance of a novel DBP that binds the target DNA more probable than any of the other modes by which the cells can escape the designed selections.

Overview: Choice of target DNA binding sequence for development of successful novel DBPs:

Our goal is the development, in part by conscious design and in part by in vivo selection, of a protein which binds to a DNA sequence of significance, e.g., a structural gene or a regulatory element, and through such binding inhibits or enhances its biological activity. In the preferred embodiment, the protein represses transcription of a deleterious element, such as a viral gene. A sufficiently long sequence could be the target of several independently acting DBPs.

Another goal of this invention is to derive one or more DBPs that bind sequence-specifically to any predetermined target DNA subsequence. It is not yet possible to design the DBP-domain amino-acid sequence from a set of rules appropriate to the target DNA subsequence. Rather, it is possible to pick sets of residues that can affect the DNA recognition of a parental DBP. Then, variegation of residues that affect DNA recognition coupled with selection for binding to the target DNA subsequence can produce a novel DBP specific for the target DNA subsequence. Such a method is limited by the number of amino acids that can be varied at one time. To develop a novel DBP that recognizes 15 bases could require changing 15 or more residues in the initial DBP. Variegation of 15 residues through all 20

amino acids would produce $20^{15} = 3.3 \times 10^{19}$ sequences and is beyond current technology. Thus we start with the recognition sequence of the initial DBP, change two to five bases and select, in one or more rounds of variegation and selection, a novel DBP that recognizes this new target DNA subsequence. This new DBP becomes the parent to the next step in which the target DNA subsequence is changed by an additional two to five bases so that a stepwise series of changes in binding protein and changes in target is used. It is emphasized here that, although we initially select DBPs that recognize sequences similar to that recognized by the IDBP, the ultimate target sequence recognized by the desired final DBP can be completely unrelated to the recognition sequence of the IDBP.

The process of finding a DBP that recognizes a sequence within a genome is shortened if we pick sequences that have some similarity to the cognate sequence of the initial DBP. The intent is to locate several unique sites in the gene which can be bound specifically by DBPs such that transcription through those sites is reduced.

The sequences of some regions of genes of eukaryotic pathogens vary among strains (SAAG88). To optimize the search for target sites in the gene selected for repression such that repression will be effective in all or the majority of strains of a pathogen, regions of conserved DNA sequence within the gene are, preferably, identified.

There may be a very small number of sequences that occur in the genome of the host cells for which binding of a DBP will be lethal. For this reason, the regulatory sequences, such as promoters, of the host organism are not preferred targets for DBP development. Preferably, the target sequence occurs only in the gene of interest. For some applications, target sequences that occur at locations other than the site f intended action may be used if binding of a protein to the extra sites is acceptable.

Preliminary elimination of non-unique sequences is done by searching DNA sequence data banks of host genomic sequences and bacterial strain sequences, and by searching the plasmid sequences for matches to the potential target subsequences. Remaining potential target subsequences are then used as oligonucleotide probes in Southern analyses of host genomic DNA and bacterial DNA. Sequences which do not anneal to host or bacterial DNA under stringent conditions are retained as target subsequences. These target subsequences are cloned into the operative vector at the promoters of the selection genes for DBP function, as described for the test DNA binding sequence.

Choice of target subsequences is based also on the optimal location of target sites within a gene such that transcription will be maximally affected. Studies of monkey L-cells show that lac repressor can bind to lac operator, or to two lac operators in tandem, in the L-cell nucleus (HUMC87, HUMC88). Further, this binding results in repression of a downstream chloramphenicol acetyl transferase gene in this system, and repression is relieved by IPTG. Two tandem operators repress CAT enzyme production to a greater extent than a single operator. The user preferably locates two to four target sites relatively close to each other within the transcriptional unit.

Overview: Strategies for Obtaining Protein Recognition of Non-Symmetric Target DNA Sequences

In vitro, lac repressor binds to a perfectly palindromic synthetic lac operator which omits the central base pair of the natural operator 10 times more tightly than it does to the wild-type operator (SADL83). In vivo, the synthetic operator represses beta-galactosidase activity to a 4-fold lower level than does the wild-type repressor. Simons et al. (SIM084) describe the isolation of five lac operator-like subsequences from eukaryotic DNA that titrate

lac repressor in vivo. All five subsequences share a 14 bp consensus subsequence that lacks the central base pair of the natural lac operator and is a perfect palindrome of the left seven base pairs of the natural lac operator. A synthetic 11-base pair inverted repeat of the left half of the E. coli lac operator binds lac repressor 8-fold more tightly than does the natural operator. We conclude that natural repressors have not evolved to have maximal affinity for their operators, rather they have evolved to produce optimal regulation.

E. coli trp repressor (BASS87) and λ repressor (BENS88) symmetrized operator subsequences bind their respective repressors more tightly than do the natural operators. For λ repressor, unlike lac repressor, the optimal binding subsequence both includes a base pair at the center of symmetry and contains a non-consensus base pair (BENS88).

It is important to note that the focus of all of the above experiments has been on symmetry: symmetric operators, symmetric changes in protein binding residues, etc. In the natural systems discussed above, increasing operator subsequence symmetry towards the consensus palindrome does indeed increase the strengths of the binding interactions. This result arises, however, not from symmetry per se, but from optimizations of the protein-DNA interactions at both operator half-sites. If the DNA-binding protein presents a different binding domain to the operator at each half-site, symmetric DNA operator subsequences are not only not optimal but are unfavorable. The implications of this distinction have not been considered in the literature.

Starting from natural, dyad symmetric or de novo designed DBPs we can generate specific DBPs with non-symmetric target recognition using a variety of strategies. Seven examples of strategies are listed; however, this invention is not limited to these particular strategies.

1) Produce two dimeric DBPs. One DBP is produced by the means described here to recognize a symmetrized version of the left half of the target and is called DBP_L. The other DBP is similarly produced to recognize a symmetrized version of the right half of the target and is called DBP_R. Cells producing equimolar mixtures of DBP_R and DBP_L contain approximately 1 part DBP_L dimer, 2 parts DBP_L:DBP_R heterodimer, and 1 part DBP_R dimer. Thus one half of the DBP molecules bind to the non-symmetric target subsequence. These heterodimers may be isolated by affinity separation techniques, or the 50% active mixture may be used directly.

2) Produce a mixture of DBP_R and DBP_L as described in (1) and crosslink proteins with an agent such as glutaraldehyde. Use a column that contains the DNA target subsequence to purify DBP_L:DBP_R heterodimer from the homodimers.

3) Produce (by variegation of the dimerization interface of a known DBP, as described more fully hereafter) a heterodimer comprised of complementing mutant sequences DBP1 and DBP2 such that the heterodimer DBP1:DBP2 is exclusively formed. Next, alter the recognition domains of DBP1 and DBP2 by the methods described here to produce heterodimers having asymmetric recognition, e.g. $DBP1_L$:$DBP2_R$.

4) Produce a heterodimer DBP1:DBP2 as in (3) and crosslink the proteins in vitro with an agent such as glutaraldehyde as in (2).

5) Produce two dimeric DBPs with left and right target recognition elements as in (1); produce complementing heterodimer mutations as in (3) such that the non-

symmetric recognition heterodimer $DBP1_L$:$DBP2_R$ is constructed.

6) Produce a pseudo-dimer composed of a single polypeptide chain such that recognition elements that contact different bases are encoded by different codons; each DNA-contacting residue and every domain is independently variable and so asymmetric recognition can be established.

7) Produce DBP_L and DBP_R in separate steps where heterodimers of DBP:DBP_R is developed to recognize a hybrid target consisting of the wild type left half-site fused to the right half of the target and DBP_L:DBP is developed to recognize a hybrid target consisting of the wild type right half-site fused to the left half of the target. Once produced, DBP_L and DBP_R are co-expressed intracellularly as described in (1) above, crosslinked as described in (2) above, or are modified to produce the obligately complementing non-symmetric recognition heterodimer DBP_L:DBP_R as described in (5) above.

Detailed Example 1 employs strategy 5; Detailed Example 2 employs strategy 6. Section 6 of Detailed Example 1 also describes strategy 3.

For each target DNA sequence chosen, a left arm T_L, a center core T_C and a right arm T_R are defined. Two symmetrized derivatives of this target subsequence, the left symmetrized target $T_L^{->}$-T_C-$T_L^{<-}$ and the right symmetrized target $T_R^{<-}$-T_C-$T_R^{->}$ are designed and synthesized.

We divide the target DNA sequence into T_L, T_C, and T_R based on knowledge of the interaction of the parental DBP with DNA sequences to which it binds, i.e. the operator. This knowledge may come from X-ray structures of parental DBP-operator complexes, models based on 3D structures of

the DBP, genetics, or chemical modification of parental DBP-operator complexes.

Our strategy is to pick a target by finding a sequence that contains a close approximation to the central core of the operator. Bases in the center of the target may not be contacted directly by the DBP but affect the specificity of binding by influencing the position or flexibility of the bases that are contacted directly by the DBP. Accommodating changes (operator vs. target) in uncontacted bases may require subtle changes in the tertiary or quaternary structure of the DBP, such as might be effected by alterations in the dimerization interface of a dimeric DBP. We can accommodate most changes in bases directly contacted by the DBP by altering the residues that contact those bases. Therefore, it is easier to accommodate changes in those bases that are directly contacted by the DBP and we endeavor to avoid changes in the central core by seeking a target the central core of which is highly similar to the central core of the operator of the parental DBP.

We must balance two tendencies: a) if we assign too many bases to T_C, we are unlikely to find a close approximation of T_C in the genome of interest; and b) if we assign too few bases to T_C, we may thereby assign uncontacted bases to the arms. Differences between the target and the DNA sequence that binds the initial DBP at uncontacted bases in the arms may be difficult to accommodate through variegation of residues that contact the DNA directly; such a situation could cause variegation and selection to yield a functional DBP very slowly. Preferably, the length of T_C is at least 6 but not greater than 10.

We search the target genome, first with the entire operator binding sequence, and then with progressively shorter central fragments of the operator, until an acceptable match is found. A match is acceptable if all or

almost all the bases (e.g. six out of seven) match and other criteria are met.

Consider matching 6 of 7 bases as the criterion for choosing a target. The original sequence is acceptable, as are the 21 (=7 x 3) sequences that differ by one base. There are $4^7 = 2^{14} = 16384$ possible heptamers. Thus we should expect to find an acceptable match every 16384/22 = 745 bases. Similarly, matching 7 of 8 bases should occur every 65536/25 = 2622 bases; matching 8 of 9 bases should occur every 262144/28 = 9362 bases. These expected frequencies are such that viruses, which have genome sizes ranging from 5×10^3 bases up to 10^6 bases or more, should have one or more matches of 6 of 7 bases. Larger viruses should contain matches of 7 of 8 or even 8 of 9 bases.

Other criteria may include restricting the search to parts of the genome not known to vary among different isolates of the organism.

If the longest matching search sequence is such that bases known to have no direct contact with the DBP are assigned to the arms, then we increase the size of T_C to at least seven and then use a progression of core sequences to move in a stepwise fashion from a sequence that closely resembles the operator of the parental DBP to that of the target. We obtain an acceptable DBP for each target by variegation and selection. The best DBP from one target becomes the parental DBP for the next target in the progression. Accommodating changes in uncontacted bases in the central core may require variegation of residues in the protein:protein interface to produce subtle changes in the tertiary and quaternary structure of the DBP.

To illustrate this process, we consider the target chosen for Detailed Example 1. The HIV 353-369 target subsequence ACTTTCCGCTGGGGACT is nucleotides 353 to 369 of the HIV-1 genome (RATN85), chosen because of the close

match of the central 7 bp of the Kim Consensus sequence (CCGCGGG) of λ Cro (KIMJ87) to the underscored bases. No non-variable HIV sequence matched the nine central bases of any O_R3-like sequence. Highly conserved bases of λ O_R3 are written bold with stars above.

```
                    123456789
                     * * *     * * *
          O_R3   5' TATCACCGCAAGGGATA 3'
                 3' ATAGTGGCGTTCCCTAT 5'

         HIV-1   5' ACTTTCCGCTGGGGACT 3'
                 3' TGAAAGGCGACCCCTGA 5'
                353 ↑
```

T_L, T_C, and T_R are defined, in this case, as the first 5 bases, the center 7 bases (underlined), and the last 5 bases, respectively, of the target subsequence. $T_L^{->}$ differs from the corresponding bases of O_R3 at four of five bases, including the strongly conserved A2 and C4. $T_L^{<-}$ is complementary to $T_L^{->}$:

$T_L^{->}$ = 5' ACTTT 3'
$T_L^{<-}$ = 3' TGAAA 5'

We create the symmetrized target by rotating $T_L^{<-}$ about the center of the 17 bp sequence into the same strand as $T_L^{->}$:

$T_L^{->}$ = 5' ACTTT AAAGT 3' = $T_L^{<-}$
$T_L^{<-}$ = 3' TGAAA_____↑ 5'

$T_R^{<-}$ differs from the corresponding bases of O_R3 at three of five positions, including the highly conserved A2. We rotate $T_R^{<-}$ into the same strand as $T_R^{->}$:

$T_R^{<-}$ = 5' AGTCC GGACT 3' = $T_R^{->}$
3' ↑_____CCTGA 5' = $T_R^{<-}$

Symmetrized operators derived from HIV 353-369 are:

Left Symmetrized Target:

5' A C T T T C C G C T G G A A A G T 3'
3' T G A A A G G C G A C C T T T C A 5'
| $T_L^{->}$ | T_C | $T_L^{<-}$ |

and

Right Symmetrized Target:

5' A G T C C C C G C T G G G G A C T 3'
3' T C A G G G G C G A C C C C T G A 5'
| $T_R^{<-}$ | T_C | $T_R^{->}$ |

The two symmetrized derivatives are engineered into the appropriate vectors so that each of these sequences regulates the expression of the designed selectable genes of each of the respective vectors.

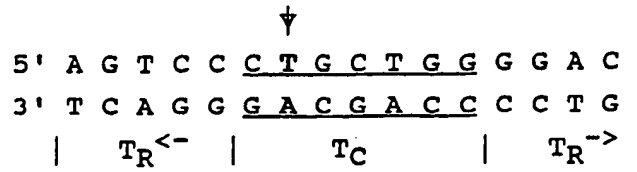
Had the best match in HIV to the CCGCGGG core been, for example, CTGCTGG, then we would use the symmetrized targets shown above until we found acceptable DBPs for the right and left targets. At that point, we would change the symmetrized targets to :

Second Left Symmetrized Target:

↓
5' A C T T T C T G C T G G A A A G T 3'
3' T G A A A G A C G A C C T T T C A 5'
| $T_L^{->}$ | T_C | $T_L^{<-}$ |

and

Second Right Symmetrized Target:



Using these targets and the selected right and left DBPs as new parental DBPs, we would initiate a new round of variegation and selection.

As another example, consider the 14 bp 434 operator. We could take each arm as 4 bp and the central core as 6 bp. We are likely to find good matches to the 6 bp core in any genome larger than 4096 bases. Thus, this division of the operator is preferred over that which assigns 5 bp to each arm and 4 bp to the core.

In order to obtain proteins that bind to these symmetrized targets, we generate a population of potential dbp genes by synthesizing DNA that codes on expression for part or all of a potential DBP and having variegated bases in the codons that encode residues of the parental DBP that are thought to contact the DNA or that influence the detailed position or dynamics of residues that contact the DNA. The variegation in the chosen codons, embodied in the synthetic DNA, is transfered to the pdbp gene either by replacement of a cassette or by annealing a mutagenic oligonucleotide to ssDNA.

The pdbp gene may be part of the vector that carries the selectable binding marker genes or may be separate. Two sets of selectable binding marker genes are prepared, one carrying the Right Symmetrized Targets (RST) and one carrying the Left Symmetrized Targets (LST). If the pdbp gene is on a different vector from the selectable binding marker genes, then RST and LST selection strains are prepared. A highly variegated population of pdbp genes is delivered into cells that also contain one of: a) the RST-

containing selectable genes, or b) the LST-containing selectable genes.

The two sets of transformed cells are selected for vector uptake and successful repression at low stringency of selection. In the case described in Detailed Example 1, cells containing DBPs will be Tc^S, Fus^R, and Gal^R.

After one or more variegation steps, DBPs that bind tightly and specifically to each of the Left Symmetrized and Right Symmetrized Targets are obtained. These DBPs are designated, in general terms, DBP_L and DBP_R, respectively. If these proteins are produced in equal amounts in the same cell, then approximately 50% of DBP protein dimers consist of the DBP_L:DBP_R heterodimer. This may be sufficient for repression of the target. In the preferred embodiment, further mutations are introduced into the DBP_L and DBP_R proteins, as described below, to enable 100% of the molecules to form heterodimers.

In an especially preferred embodiment, variegation of the gene to alter its DNA-specificity is combined with variegation of the gene to alter its dimerization (protein-protein binding) characteristics, so that the formation of the heterodimer DBP_L:DBP_R is favored. The variegation of the dimerization interface may precede (strategy 3) or follow (strategy 5) the alteration of the DNA specificity. Simultaneous variegation at both sites is also possible.

The DNA-binding proteins considered here interact with specific DNA sequences as multimers (usually dimers or tetramers) (PABO84). Monomers usually associate independently and the resulting multimer interacts with DNA. Coupling between oligomerization and DNA-binding equilibria results in explicit inclusion of oligomerization effects in the apparent affinity of DNA-binding proteins for their operators (JOHN80, RIGG70, and CHAD71).

The precise geometry of the protein in the complex with DNA strongly influences the strength of the interaction with DNA. For example, Sauer et al. generated a 92 amino-acid fragment of λ repressor carrying the YC88 mutation. This N-terminal domain dimerizes through a covalent S-S bond. Although the dissociation into monomers is unmeasurable, the binding to DNA is diminished about 10-fold relative to intact λ repressor (SAUE86).

The results presented by Reidhaar-Olson and Sauer (REID88), summarized in Table 7, show which residues in the dimerization region, when varied, will produce functional homodimers of N-terminal domains with little alteration of structure. Wide variation is tolerated at solvent exposed positions 85, 86, and 89. In contrast, almost no substitutions are tolerated at the buried positions 84 and 87. Most hydrophobic residues are functional at position 91 (except P (We use the single-letter code for amino acids, AUSU87, Appendix A)) although aromatic residues are excluded. The hydrophobic interactions among I84, M87' and V91' had previously been shown to be major components of dimerization free energy (NELS83, WEIS87b). In general, mutations that destabilize λ repressor N-terminal dimerization are similar to those that destabilize global protein structure.

The P22 Mnt repressor, like λ Cro, is a small protein containing both DNA-binding and oligomerization sites. Unlike Cro, P22 Mnt is a tetramer in solution (VERS85b, VERS87a). The amino acid sequence of Mnt has been determined (VERS87a) but the three dimensional structure of the protein is not known. Knight and Sauer (KNIG88) have shown, by sequential deletion of C-terminal residues, that Y78 is essential for tetramer formation.

A preferred embodiment of this process utilizes information available on protein structure obtained from crystallographic, modeling, and genetic sources to predict

the residues at which mutation results in stable protein monomers that retain substantially the same 3D structure as the wild-type DBP, but that fail to form dimers. Dimerization mutants are constructed using site-directed mutagenesis to isolate one or more user specified substitutions at chosen residues. The process starts using one of the genes selected for binding to a symmetrized target, denoted $\underline{dbp}_1$ ($\underline{dbp}_1$ could be either the $\underline{dbp}_L$ gene or the $\underline{dbp}_R$ gene), as the parental sequence, so that each of several specific mutations is engineered into the gene for a protein binding specifically to the symmetrized target used in the selection (the Left Symmetrized Target in the case of the $\underline{dbp}_L$ gene).

Reverse selection isolates cells not expressing a protein that binds to the target DNA sequence. This phenotype could arise in several ways, including: a) a mutation or deletion in the $\underline{dbp}_1$ gene so that no protein is produced, b) a mutation that renders the descendant of the parental DBP_1 unstable, c) a mutation that allows the descendant of the parental DBP_1 to persist and to fold into nearly the same 3D structure as the parental DBP, but which prevents oligomerization. It is anticipated that reverse selection will isolate many genes for non-functional proteins and that these proteins must be analyzed until a suitable oligomerization-mutant is found. Therefore, we choose sites carefully so that we maximize the chance of disrupting oligomerization without destroying tertiary structure. We also use lower levels of variegation in reverse selection so that the number of mutants to be analyzed is not too large. For forward selection, the number of different mutants is preferably 10^4 to 10^9, and more preferably greater than 10^6. For reverse selection, it is 10^3 to 10^6. (Under certain circumstances, the number of reverse selection mutants could be as low as 10-20).

Cassettes bearing the site-specific changes are synthesized and each is ligated into the vector at the

appropriate site in the $\underline{dbp_1}$ gene. Transformants are obtained by the antibiotic-resistance selection for vector maintenance (e.g. Ap), and screened for loss of repression of the selective systems under control of Dbp_1 binding. Defective dimerization results in substantially decreased DNA affinity, hence the altered derivatives are recognized by screening isolates obtained using the selectable gene systems. In Detailed Example 1 (where $\underline{dbp_1}$ is $\underline{dbp_L}$), dimerization-defective derivatives are Tc^R, Fus^S and Gal^S in E. coli delta4 cells (Gal^+ in cells of E. coli strain HB101). Appropriate controls are used to verify that the loss of repression is due to a substitution in $\underline{dbp_L}$.

In Detailed Example 1 (using an engineered synthetic λ cro gene designated rav), the mutant Rav protein with specific binding to the Left Symmetrized Target (designated Rav_L; gene, $\underline{rav_L}$) is used to produce a derivative defective in dimerization. Studies of λ Cro suggest that the dimer is stabilized by interactions in an antiparallel beta sheet between residues E54, V55 and K56 from each monomer (ANDE81, PABO84). In addition, F58 appears to stabilize the Cro dimer through hydrophobic interactions between F58 of one monomer and residues in the hydrophobic core of the other monomer (TAKE85). Further, mutational studies (PAKU86) show that some substitutions at E54 and at F58 result in decreased intracellular specific protein levels and that these mutant proteins lack repressor activity. Mutants are constructed by using site specific mutagenesis to isolate VF55 and FW58 mutants of Rav_L. (Point mutations are written as XYnnn, where X is the amino acid found at location nnn and Y is the amino acid found in the mutant.) The cassettes bearing mutations that confer the VF55 and FW58 substitutions are synthesized, and each is ligated into the operative vector at the appropriate site within the $\underline{rav_L}$ gene. Selections and characterizations are as described above. These alleles are designated $\underline{rav_L}$-55 and $\underline{rav_L}$-58.

Alternative methods of obtaining dimerization-defective DBP derivatives are not excluded. Thus the rav$^+$ gene (Detailed Example 1) or a potential-dbp$^+$ gene coding for any globular dimerizing protein, may be subjected to Structure-directed Mutagenesis of residues involved in the protein-protein dimer interface. In the case of the rav$^+$ allele, residues 7, 23, 25, 30, 33, 40, 42, 52, 54, 55 and 58 are candidates for mutagenesis.

For example, mutagenesis of rav$^+$ residues 52, 54, 55 and 58, using a cassette carrying vgDNA at codons specifying these residues, is followed by ligation and transformation of cells. Selection is applied for plasmid maintenance (Ap^R) and loss of repression (Tc^R, and galactose utilization in HB101 cells). Variegated plasmid DNA is purified from a population of Ap^R Tc^R Gal$^+$ cells and analyzed with restriction enzymes and Southern blotting. Plasmid preparations containing the vg-rav fragment of predominantly rav$^+$ molecular weight are retained, and are designated vg-ravA.

To isolate a second dimer-specific rav mutant protein, designated RavB, such that the mutation in ravB is complementary to a mutation contained in the vg-ravA population, Structure-directed Mutagenesis is performed on a second copy of the rav gene, designated ravB, carried on a plasmid conferring a different antibiotic resistance (e.g. Km^R). Residues affecting the same dimer interface are varied. Competent vg-ravA cells are transformed with the vg ravB plasmid preparation. Transformants are obtained as Ap^R Km^R, and further selected for Rav$^+$ phenotype using the selection systems (Tc^S, Fus^R, Gal^R in an E. coli delta4 cell genetic background).

The surviving colonies are analyzed by restriction analysis of plasmids, and are backcrossed to obtain pure plasmid lines that confer each of the Ap^R and Km^R phenotypes. In this manner, mutants bearing obligate comple-

menting dimerization alleles of ravA and ravB are isolated. These rav mutations may be tested pairwise to confirm complementation, and are sequenced. The information obtained from these mutants is used to introduce these dimerization mutations into rav_L and ravR genes previously altered by Structure-directed Mutagenesis in DNA-binding specificity domains as described above.

In one preferred embodiment of this invention (strategy 3), isolation of dbp_L and dbp_R mutations that confer specific and tight binding to target DNA sequences $T_L^{->}$-T_C-$T_L^{<-}$ and $T_R^{<-}$-T_C-$T_R^{->}$ is followed by engineering of second site mutations causing a dimerization defect, for example dbp_L-1 as described herein. Complementing mutations are introduced into each of the dbp_R and dbp_L genes, such that obligate heterodimers are co-synthesized and folded together in the same cell and bind specifically to the non-palindromic targets.

A primary set of residues is identified. These residues are predicted, on the basis of crystallographic, modeling, and genetic information, to make contacts in the dimer with the residue altered to produce DBP_L-1. A secondary set of residues is chosen, whose members are believed to touch or influence the residues of the primary set. An initial set of residues for Focused Mutagenesis in the first variegation step is selected from residues in the primary set. A variegation scheme, consistent with the constraints described herein, is picked for these residues so that the chemical properties of residues produced at each variegated codon are similar to those of the wild-type residue; e.g. hydrophobic residues go to hydrophobic or neutral, charged residues go to charged or hydrophilic. A cassette containing the vgDNA at the specified codons is synthesized and ligated into the dbp_R gene carried in a vector with a different antibiotic selection than that on the vector carrying the dbp_L-1 gene. For example, in Detailed Example 1, rav_L-55 or rav_L-58 are encoded on

plasmids that carry the gene for Ap^R. Variegated $\underline{rav}_R$ genes are cloned into plasmids bearing the gene for Km^R.

The protein produced by the $\underline{dbp}_L$-1 allele carrying the dimerization mutation fails to bind to the $\underline{dbp}_L$ target $T_L^{->}$-T_C-$T_L^{<-}$. Cells bearing this target as a regulatory site upstream of selection genes display the DBP^- phenotype. This phenotype is employed to select complementary mutations in a $\underline{dbp}_R$ gene. Following ligation of the mutagenized cassette into the appropriate (e.g. Km^R) plasmid, cells bearing $\underline{dbp}_L$-1 on a differently marked (e.g. Ap^R) plasmid are made competent and transformed. Transformants that have maintained the resident plasmid (e.g. Ap^R Km^R) are further selected for DBP^+ phenotype in which binding of the non-palindromic target subsequence $T_L^{->}$-T_C-$T_R^{->}$ is required for repression. Only heterodimers consisting of a 1:1 complex of the $\underline{dbp}_L$-1 gene product and the complementing $\underline{dbp}_R$-1 gene product bind to the target, and produce Tc^S Fus^R Gal^R colonies in the appropriate cell host.

Each resident plasmid is obtained from candidate colonies by plasmid preparation and transformation at low plasmid concentration. Strains carrying plasmids encoding either mutant $\underline{dbp}_L$-1 or mutant $\underline{dbp}_R$-1 genes are selected by the appropriate antibiotic resistance (e.g. in Detailed Example 1, Ap^R or Km^R selection, respectively). Plasmids are independently screened for the DBP^- phenotype, characterized by restriction digestion and agarose gel electrophoresis, and plasmid pairs are co-tested for complementation by restoration of the DBP^+ phenotype with respect to the $T_L^{->}$-T_C-$T_R^{->}$ target when both $\underline{dbp}$ alleles are present intracellularly. Successfully complementing pairs of $\underline{dbp}$ genes are sequenced. Subsequent variegation steps may be required to optimize dimer interactions or DNA binding by the heterodimer.

In Detailed Example 1, the VF55 change in Rav_L-55 introduces a bulky hydrophobic side group in place of a smaller hydrophobic residue. A complementary mutation inserts a very small side chain, such as G55 or A55, in a second copy of the protein. In this case, the primary set for mutagenesis is V55. A secondary set of residues includes nearby components of the beta strand E54, K56, P57, and E53 as well as other residues. In the initial variegation step, residues 53-57 are subjected to Focused Mutagenesis, such that all amino acids are tested at this location. Cells containing complementing mutant proteins are selected by requiring repression of the nonpalindromic HIV 353-369 target subsequence $T_L^{->}$-T_C-$T_R^{->}$.

In another embodiment, the $\underline{rav}_L$-58 allele carrying the substitution FW58 and conferring the Rav^- phenotype, is used for selection of complementing mutations following Structure-directed Mutagenesis of the $\underline{rav}_R$ gene. Residues L7, L23, V25, A33, I40, L42, A52, and G54 are identified as the principal set. Residues in the secondary set include F58, P57, P59, and buried residues in alpha helix 1. In the initial variegation step residues 23, 25, 33, 40, and 42 are varied through all twenty amino acids. Subsequent iterations, if needed, include other residues of the primary or secondary set. In this manner, $\underline{rav}_R$-1, -2, -3, etc. are isolated, each of which yields a protein that is an obligate complement of the $\underline{rav}_L$-58 mutation. Selection for Rav^+ phenotype using the HIV 353-369 target $T_L^{->}$-T_C-$T_R^{->}$ sequence is used as described in the preferred embodiment.

In either the preferred or alternative embodiments, this process teaches a method of constructing obligate complementing mutations at an oligomer interface. These pairs of mutations may be used in further embodiments to engineer novel DBPs specific for HIV 681-697 and HIV 760-776 targets; for targets in other pathogenic retroviruses such as HTLV-II; for other viral DNA-containing pathogens

such as HSVI and HSVII; as well as for non-viral targets such as deleterious human genes. Similar methodology is claimed for engineering DBPs for use in animal and plant systems.

Overview: Selection of the Initial DNA-Binding Protein for Variegation

The choice of an initial DBP is determined by the degree of specificity required in the intended use of the successful DBP and by the availability of known DBPs. The present invention describes three broad alternatives for producing DBPs having high specificity and tight binding to target DNA sequences. The present invention is not limited to these classes of initial potential DBPs.

A first alternative is to use a polypeptide that will conform to the DNA and can wind around the DNA and contact the edges of the base pairs. A second alternative is to use a globular protein (such as a dimeric H-T-H protein) that can contact one face of DNA in one or more places to achieve the desired affinity and specificity. A third alternative is to use a series of flexibly linked small globular domains that can make contact with several successive patches on the DNA.

DNA features influencing choice of an initial DBP:

Features of DNA that influence the choice of an initial DBP include sequence-specific DNA structure and the size of the genome within which the DBP is expected to recognize and affect gene expression.

Sequence-specific aspects of DNA structure that can influence protein binding include: a) the edges of the bases exposed in the major groove, b) the edges of the bases exposed in the minor groove, c) the equilibrium positions of the phosphate and deoxyribose groups, d) the

flexibility of the DNA toward deformation, and e) the ability of the DNA to accept intercalated molecules. Note that the sequence-specific aspects of DNA are carried mostly inside a highly charged molecular framework that is nearly independent of sequence.

The strongest signals of sequence are found in the edges of the base pairs in the major groove, followed by the edges in the minor groove. The groove dimensions depend on local DNA sequence (NEID87b, KOUD87, ULAN87).

The number of base pairs required to define a unique site depends on the size and non-randomness of the genome. Consider a genome of length Z_g bases and consider a specific subsequence of length Q. If the genome is random, the subsequence is expected to occur N(Q) times, where

$$N(Q) = \frac{2\ Z_g}{4^Q} = \frac{2\ Z_g}{2^{2Q}} .$$

From this equation, we derive the expression Q_u, which is the lower limit of the length of subsequences that are expected to occur once or be absent:

$$Q_u = \log_2(2\ Z_g)/2.$$

Z_g	$\log_2 (2\ Z_g)/2$	Q_u
10^6	10.5	11
10^7	12.1	13
10^8	13.8	14
10^9	15.5	16
10^{10}	17.1	18

Thus, a DNA subsequence comprising 12 base pairs may be unique in the E. coli genome (5×10^6 bp), but is likely to

occur about 180 times in a random sequence the size of the human genome (3×10^9 bp).

The non-random nature of DNA sequences in genomes has been shown to result in the over- and under-representation of specific sequences. The random-genome model can underestimate the probe length needed to define a unique coding sequence (LATH85). Recognition sites for certain restriction enzymes occur in clusters and are found much more often than expected (SMIT87). In contrast, lac repressor binding sites in eukaryotic genomes are almost two orders of magnitude less frequent than expected on the basis of random sequence (SIM084).

Protein features influencing choice of initial DBP:

Sequence-specific binding to DNA by DBPs does not require unpairing of the bases. Most sequence-specific binding by proteins to DNA is thought to involve contacts in the DNA major groove.

To be certain of unique recognition in the human genome, it is best to design a protein that recognizes 19 to 21 base pairs. To contact 20 base pairs directly, a protein would need to: a) wind two full turns around the DNA making major groove contacts, b) make a combination of major groove and minor groove contacts, or c) contact the major groove at four or five places. An extended polypeptide, binding in the major groove of B-DNA, lies about 5.0 A from the DNA axis. One base pair and 1 1/2 amino acids extend roughly equal distances along the helix (SAEN83, p238).

A nine residue alpha helix, such as the recognition helices of H-T-H repressors, extends about 13.5 Å along the major groove. If residues with long side chains are located at each terminus of the helix, the helix can make contacts over a 20.0 Å stretch of the major groove allowing

six base pairs to be contacted. Parts of the DBP other than the second helix of the H-T-H motif can make additional protein-DNA contacts, adding to specificity and affinity. The rigidity of the alpha helix prevents a long helix from following the major groove around the DNA. A series of small domains, appropriately linked, could wind around DNA, as has been suggested for the zinc-finger proteins (BERG88a, GIBS88, FRAN88). In an extended configuration a polypeptide chain progresses roughly 3.2 to 3.5 Å between consecutive residues. Thus, a 10 residue extended protein structure could contact 5 to 8 bases of DNA.

Stable complexes of proteins with other macromolecules involve burial of 1000 $Å^2$ to 3000 $Å^2$ of surface area on each molecule. For a globular protein to make a stable complex with DNA, the protein must have substantial surface that is already complementary to the DNA surface or can be deformed to fit the surface without loss of much free energy. Considering these modalities we assign each genetically encoded polypeptide to one of three classes:

1) a polypeptide that can easily deform to complement the shape of DNA,

2) a globular protein, the internal structure of which supports recognition elements to create a surface complementary to a particular DNA subsequence, and

3) a sequential chain of globular domains, each domain being more or less rigid and complementary to a portion of the surface of a DNA subsequence and the domains being linked by amino acid subsequences that allow the domains to wind around the DNA.

Compl mentary charges can accelerate association of molecules, but they usually do not provide much of the free energy of binding. Major components of binding energy arise

from highly complementary surfaces and the liberation of ordered water on the macromolecular surfaces.

Properties of sequence-specific DNA-binding by polypeptides:

An extended polypeptide of 24 amino acids lying in the major groove of B-DNA could make sequence-specific interactions with as many as 15 base pairs, which is about the least recognition that would be useful in eukaryotic systems. Peptides longer than 24 amino acids can contact more base pairs and thus provide greater specificity.

Extended polypeptide segments of proteins bind to DNA in natural systems (e.g. λ repressor and Cro, P22 Arc and Mnt repressors). The DNA major groove can accommodate polypeptides in either helical or extended conformation. Side groups of polypeptides that lie in the major groove can make sequence-specific or sequence-independent contacts. Since the polypeptide can lie entirely within the major groove, contacts with the phosphates are allowed but not mandatory. Thus a polypeptide need not be highly positively charged. A neutral or slightly positively charged polypeptide might have very low non-specific binding.

Polypeptides composed of the 20 standard amino acids are not flat enough to lie in the minor groove unless the sequence contains an extraordinary number of glycines, however, residue side-groups could extend into the minor groove to make sequence-specific contacts. Polypeptides of more than 50 amino acids may fold into stable 3D structures. Unless part of the surface of the structure is complementary to the surface of the target DNA subsequence, formation of the 3D structure competes with DNA binding. Thus polypeptides generated for selection of specific binding are preferably 25 to 50 amino acids in length.

Polypeptides present the following potential advantages:

a) low molecular weight: an extended polypeptide offers the maximum recognition per amino acid,

b) polypeptides have no inherent dyad symmetry and so are not biased toward recognition of palindromic sequences,

c) polypeptides may have greater specificity than globular proteins, and

d) peptides may be good models from which other low molecular weight compounds may be designed.

Thus, one would choose a polypeptide as initial DNA-binding molecule if high specificity and low molecular weight are desired.

No sequence-specific DNA-binding by small polypeptides has been reported to date. Possible reasons that such polypeptides have not been found include: a) no one has sought them, b) cells degrade polypeptides that are free in the cytoplasm, and c) they are too flexible and are not specific enough.

In a preferred embodiment, a DNA-binding polypeptide is associated with a custodial domain to protect it from degradation, as discussed more fully in Examples 3 and 4.

<u>Properties of globular proteins influencing choice of initial DBP:</u>

The majority of the well-characterized DBPs are small globular proteins containing one or more DNA-binding domains. No single-domain globular protein comprising 200 or fewer amino acids is likely to fold into a stable

structure that follows either groove of DNA continuously for 10 bases. The structure of a small globular protein can be arranged to hold more than one set of recognition elements in appropriate positions to contact several sites along the DNA thereby achieving high specificity, however, the bases contacted are not necessarily sequential on the DNA. For example, each monomer of λ repressor contains two sequence-specific DNA recognition regions: the recognition helix of the H-T-H region contacts the front face of the DNA binding site and the N-terminal arm contacts the back face. To obtain tight binding, a globular protein must contact not only the base-pair edges, but also the DNA backbone making sequence-independent contacts. These sequence-independent contacts give rise to a certain sequence-independent affinity of the protein for DNA. The bases that intervene between segments that are directly contacted influence the position and flexibility of the contacted bases. If the DNA-protein complex involves twisting or bending the DNA (e.g. 434 repressor-DNA complex), non-contacted bases can influence binding through their effects on the rigidity of the target DNA sequence.

The phage repressors Arc, Mnt, λ repressor and Cro are proposed to bind to DNA at least partly via binding of extended segments of polypeptide chain. The N-terminal arm of λ repressor makes sequence-specific contacts with bases in the major groove on the back side of the binding site. The C-terminal "tail" of λ Cro is proposed to make sequence-independent contacts in the minor groove of the DNA. The structure of neither Arc nor Mnt has been determined; however, the sequence specificity of the N-terminal arm of Arc can be transferred to Mnt; viz. when Arc residues 1-9 are fused to Mnt residues 7 through the C-terminal, the fusion protein recognized the arc operator but not the mnt operator. Residues 2, 3, 4, 5, 8, and 10 of Arc have been proposed to contact operator DNA and residue 6 of Mnt has been shown to be involved in sequence-specific operator contacts.

Binding to non-palindromic sequences requires alteration of dyad-symmetric proteins. Even non-palindromic DNA has approximate dyad symmetry in the deoxyribophosphate backbone; proteins that are heterodimers or pseudo-dimers engineered from known globular DBPs are good candidates for the mutation process described here to obtain globular proteins that bind non-palindromic DNA. It has been observed that the DNA restriction enzymes having palindromic recognition are composed of dyad symmetric multimers (MCCL86), while restriction enzymes and other DNA-modifying enzymes (e.g. Xis of phage λ) having asymmetric recognition are comprised of a single polypeptide chain or an asymmetric aggregate (RICH88). Such proteins may also provide reasonable starting points to generate DBPs recognizing non-palindromic sequences.

A globular protein can bind sequence-specifically to DNA through one set of residues and activate transcription from an adjacent gene through a different set of residues (for example, λ or P22 repressors). The internal structure of the protein establishes the appropriate geometric relationship between these two sets of residues. Globular proteins may also bind particular small molecules, effectors, in such a way that the affinity of the protein for its specific DNA recognition subsequence is a function of the concentration of the particular small molecules (e.g. CRP and [cAMP]). Conditional DNA-binding and gene activation are most easily obtained by engineering changes into known globular DBPs.

Some DBPs from bacteria and bacteriophage have been shown to have sufficient specificity to operate in mammalian cells.

An initial DBP may be chosen from natural globular DBPs of any cell type. The natural DBP is preferably small so that genetic engineering is facile. Preferably,

the 3D structure of the natural DBP is known; this can be determined from X-ray diffraction, NMR, genetic and biochemical studies. Preferably, the residues in the natural DBP that contact DNA are known. Preferably the residues that are involved in multimer contacts are known. Preferably the natural operator of the natural DBP is known. More preferably, mutants of the natural operator are known and the effects of these mutants on binding by natural DBP and mutant DBPs are known. Preferably, mutations of the DBP are known and the effects on protein folding, multimer formation, and in vivo half life-time are known. Most of the above data are available for λ Cro, λ repressor and fragments of λ repressor, 434 repressor and Cro proteins, E. coli CRP and trp repressor, P22 Arc, and P22 Mnt.

Globular DBPs are the best understood DBPs. In many cases, globular DBPs are capable of sufficient specificity and affinity for the target DNA sequence. Thus globular DBPs are the most preferred candidates for initial DBP. Table 8 contains a list of some preferred globular DBPs for use as initial DBPs.

λ repressor and phage 434 repressor have been extensively studied (CHAD71, PTAS80, PABO79, JOHN79, SAUE79, SAUE86, PABO82a,b, LEWI83, OHLE83, WEIS87a,b,c, REID88, ANDE87, NELS86, ELIA85). Both proteins comprise an amino-terminal DNA-binding domein having four homologous alpha helices. Helices 2 and 3 form the H-T-H motif. DNA contacts originate in helix 2, helix 3, and adjacent regions with helix 3 providing most of the contacts. The N-terminal domains of λ repressor contact each other along helix 5 (PABO82b) while in 434 repressor the interdomain contacts are beyond helix 4, there being no helix 5 (ANDE87).

The operator DNA bends symmetrically in the 434 repressor-consensus operator co-crystal (ANDE87). The

center of the 14 base pair DNA helix is over-wound and bends slightly along its axis such that it curls around the alpha 3 helix of each repressor monomer; the ends of the operator DNA helix are underwound. Bending of operator DNA has also been proposed in models of Cro protein and CAP protein operator binding (OHLE83, GART88). Consistent with the results of Gartenberg and Crothers, bending of the 434 operator toward Cro is toward the minor groove and occurs most readily when the central bases consist exclusively of A and T (KOUD87); in this case, substitution of CG base pairs greatly reduces binding.

λ Cro (TAKE77) has been described from an X-ray structure of the protein without DNA (ANDE81). Alpha helix 2 lies across the operator major groove and may make contacts to operator backbone phosphates at its N-terminal and C-terminal ends. In addition, backbone phosphates may be contacted by residues at the C terminus of alpha 3, N terminus of beta 2, and C terminus of beta 3 (PABO84). In computer model building of λ Cro-operator DNA interactions, bending of operator DNA or bending at the monomer-monomer interface of the Cro dimer have been proposed to make the best fit between operator and dimer (PABO84).

Key amino acids within the H-T-H region of 434 Cro and λ Cro are highly conserved (PABO84), and 434 Cro binds operator DNA as a dimer (WHAR85a). Because the crystals of 434 Cro and DNA do not diffract to high resolution, atomic details of the protein-DNA interactions are not revealed (WOLB88). Nevertheless, Wolberger et al. report very significant similarities and differences between the DNA binding patterns of 434 repressor and 434 Cro. These observations on DBPs from 434, together with recent results on Trp repressor (OTWI88), support the view that a) structural elements that fit into the major groove of DNA can function in a variety of closely related ways, b) bending of DNA complexed to proteins is an important

determinant of specificity, and c) that mechanisms of recognition may be quite subtle.

Crystal structures have been determined for two DBPs, CRP (WEBE87a) and TrpR (OTWI88) from E. coli. Both these proteins contain H-T-H motifs and bind their cognate operators only when particular effector molecules are bound to the protein, cAMP for CRP and L-tryptophan for TrpR. Binding of each effector molecule causes a conformational change in the protein that brings the DNA-recognizing elements into correct orientation for strong, sequence-specific binding to DNA (JOHN86). The DNA-binding function of Lac repressor is also modulated through protein binding of an effector molecule (e.g. lactose); unlike CRP and TrpR, Lac repressor binds DNA only in the absence of the effector. CRP can act either as an activator (RENY88) or as a repressor (POLA88) depending on the relationship between the CRP-binding site and the rest of the promoter.

Two structures of CRP (MCKA81, MCKA82) and one structure of a CRP mutant (WEBE87a) are available. Otwinowski et al. (OTWI88) have published an X-ray crystal structure of TrpR bound to the Trp operator. This structure shows that, although TrpR contains a canonical H-T-H motif, the positioning of the recognition helix with respect to the DNA is quite different from the positioning of the corresponding helix in other H-T-H DBPs (MATT88) for which structures of protein-DNA complexes are available. Unlike previously determined structures, most of the interactions between atoms of TrpR and bases are mediated by localized water molecules. It is not possible to distinguish between localized water and atomic ions, such as Na^+, by X-ray diffraction alone. We shall follow Otwinowski et al. and refer to these peaks in electron density as water, although ions cannot be ruled out.

Bass et al. (BASS88) studied the binding of wild type TrpR and single amino acid missense mutants of TrpR to a

consensus palindromic Trp operator and to palindromic operators that differ from the consensus by a symmetric substitution at one base in each half operator. Bass et al. conclude that the contact between the H-T-H motif of TrpR and the operators must be substantially different from the model that had been built based on the 434 Cro-DNA structure.

Thus the binding of globular DBPs that are modulated by effector molecules is fundamentally the same as the binding of unmodulated globular DBPs, but the details of each protein's interactions with DNA are quite different. Prediction of which amino acids will produce strong specific binding is beyond the capabilities of current theory. Given the important role of localized waters or ions in the TrpR-DNA interface (OTWI88) and in the 434R-DNA interface (AGGA88), such predictions are likely to remain beyond reach for some time.

The Mnt repressor of P22 is an 82 residue protein that binds as a tetramer to an approximately palindromic 17 base pair operator presumably in a manner that is two-fold rotationally symmetric. Although the Mnt protein is 40% alpha helical and has some homology to λ Cro protein, Mnt is known to contact operator DNA by N-terminal residues (VERS87a) and possibly by a residue (K79) close to the C terminus (KNIG88). It is unlikely, therefore, that an H-T-H structure in Mnt mediates DNA binding (VERS87a). Another residue (Y78) close to the C-terminal end has been found to stabilize tetramer formation (KNIG88). Though the three dimensional structure of Mnt is not known, DNA-binding experiments have indicated that the Mnt operator, in B-form conformation, is contacted at major groove nucleotides on both front and back sides of the operator helix (VERS87a).

The Arc repressor of P22 is a 53 residue protein that binds as a dimer to a partially palindromic 21 base pair

operator adjacent to the mnt operator in P22 and protects a region of the operator that is only partially symmetric relative to the symmetric sequences in the operator (VERS87b). Arc is 40% homologous to the N-terminal portion of Mnt, and the N-terminal residues of the Arc protein contact operator DNA such that an H-T-H binding motif is unlikely, as in Mnt binding (VERS86b). The three dimensional structure of Arc, like Mnt, is not known, but a crystallographic study is in progress (JORD85). DNA-binding experiments have shown that Arc probably binds along one face of B-form operator DNA. These experiments indicate that Arc contacts operator phosphates farther out from the center of operator symmetry than do the repressors or Cro proteins of λ or 434, or P22 Mnt protein. Thus the researchers state that the operator DNA may be bent around Arc in binding or Arc dimer may have an extended structure to allow such contacts to occur (VERS87b). These alternatives are not mutually exclusive.

DNA-Binding Proteins Other Than Repressor Proteins

Any protein (or polypeptide) which binds DNA may be used as an initial DNA-binding protein; the present method is not limited to repressor proteins, but rather includes other regulatory proteins as well as DNA-binding enzymes such as polymerases and nucleases.

Derivatives of restriction enzymes may be used as initial DBPs. All known restriction enzymes recognize eight or fewer base pairs and cut genomic DNA at many places. Expression of a functional restriction enzyme at high levels is lethal unless the corresponding sequence-specific DNA-modifying enzyme is also expressed. EcoRI that lacks residues 1-29, denoted EcoRI-delN29, has no nuclease activity (JENJ86); EcoRI-delN29 binds sequence-specifically to DNA that includes the EcoRI recognition sequence, GAATTC, (BECK88).

From the structure of R.EcoRI (MCCL86), we can see that extension of the polypeptide chain at either the amino or carboxy terminus would allow contacts with base pairs outside of the canonical hexanucleotide.

Specifically, extending EcoRI(AT139), EcoRI(GS140), or EcoRI(RQ203) (YANO87) by, for example, ten highly variegated residues at the amino terminus and selecting for binding to a target such as, TGAATTCA or GGAATTCC, allows isolation of a protein having novel DNA-recognition properties. Alternatively, EcoRI may be extended at the amino terminus by addition of a zinc-finger domain. It may be useful to have two or more tandem repeats of the octanucleotide target placed in or near the promoter region of the selectable gene. Fox (FOXK88) has used DNase-I to footprint EcoRI bound to DNA and reports that 15 bp are protected. Thus, repeated octanucleotide targets for proteins derived from EcoRI should be separated by eight or more base pairs; one could place one copy of the target upstream of the -35 region and one copy downstream of the -10 region. There are many residues in EcoRI that contact the DNA as the enzyme wraps around it. These residues could be varied to alter the binding of the protein. To obtain acceptable specificity, we may need to pick as initial DBP a mutant of EcoRI that folds and dimerizes, but that binds DNA weakly. The mutations in regions of the protein that contact DNA outside of the original GAATTC will confer the desired affinity and specificity on the novel protein.

One may wish to obtain a protein that binds to one target DNA sequence, but not to other sequences that contain a subsequence of the target. For example, we may seek a protein that recognizes TGAATTCA, but not any of the sequences vGAATTCb. To achieve this distinction, we place the target sequence in the promoter region of the selectable gene and one or more instances of the related sequences, to which we intend that the protein not bind, in

the promoter region of an essential gene, such as an antibiotic-resistance gene.

Other stable proteins may also be used as initial DBPs, even if they show no DNA-binding properties. Parraga et al. (Reference 8 in PARR88) report that Eisen et al. have fused 229 residues of yeast ADR1 to beta-galactosidase and that the fusion protein binds sequence-specifically to DNA in vitro.

Adenovirus E1A protein turns on early viral genes as well as the human heat shock protein hsp70 (SIMO88). Further, a normal inducible nuclear DNA-binding protein regulates the IL-2alpha interleukin-2 receptor-R(alpha) gene and also promotes activation of transcription from the HIV-1 virus LTR (BOHN88). These studies indicate one of the many difficulties of designing antiviral chemotherapy by using the transcriptional regulatory apparatus of the virus as a target. This invention uses unique target sequences, not represented elsewhere in the host genome, as targets for suppression of gene expression.

The DNA sequences of operators that interact with proteins that control mating-type and cell-type specific transcription in yeast (MILL85) reveal that the consensus site for action of the alpha2 protein dimer is symmetric, while a heterodimeric complex of alpha2 and a1 subunits acts on an asymmetric site. The alpha2a1-responsive site consists of a half-site that is identical to the alpha2 half-site, and another half-site that is a consensus for a1 protein binding. The spacings between the symmetric and asymmetric sites are not the same.

Antibodies that bind DNA and other nucleic acids have been obtained from human patients suffering from Systemic Lupus Erythematosus. Murine monoclonal antibodies have been obtained that specifically recognize Z-DNA, B-DNA, ssDNA, triplex DNA, and certain repeating sequences

(ANDE88). Anderson et al. (ANDE88) report that: 1) the antibodies studied contact six base pairs and four phosphates, 2) antibodies are unlikely to provide some of the well known motifs for DNA-binding, e.g. helix-turn-helix, 3) study of DNA-antibody complexes may yield insights into mechanisms of recognition, and 4) a DNA-recognizing antibody might be converted into a sequence or structure specific nuclease. The shortness of the contact makes it unlikely that high specificity can be attained.

Properties of serially-linked globular domains:

A protein motif for DNA binding, present in some eukaryotic transcription factors, is the zinc finger in which zinc coordinately binds cysteine and histidine residues to form a conserved structure that is able to bind DNA (FRAN88). Xenopus laevis transcription factor TFIIIA is the first protein demonstrated to use this motif for DNA binding, but other proteins such as human transcription factor SP1, yeast transcription activation factor GAL4, and estrogen receptor protein have been shown to require zinc for DNA binding in vitro (EVAN88). Other mammalian and avian steroid hormone receptors and the adenovirus E1A protein, that bind DNA at specific sites, contain cysteine-rich regions which may form metal chelating loops.

Zinc-finger regions have been observed in the sequences of a number of eukaryotic DBPs, but no high-resolution 3D structure of a Zn-finger protein is yet available. A variety of models have been proposed for the binding of zinc-finger proteins to DNA (FAIR86, PARR88, BERG88, GIBS88). Model building suggests which residues in the Zn-fingers contact the DNA and these would provide the primary set of residues for variation. Berg (BERG88) and Gibson et al. (GIBS88) have presented models having many similarities but also some significant differences. Both models suggest that the motif comprises an antiparallel beta structure followed by an alpha helix and that the front side of the

helix contacts the major groove of the DNA. By assuming that conserved basic residues of the Zn-finger make contact with phosphate groups in each copy of the motif, Gibson et al. deduce that the amino terminal part of the helix makes direct contact to the DNA. The Gibson model does not, however, account well for the number of bases contacted by Zn-finger proteins. The observations on H-T-H proteins suggest that a DNA-recognizing element can interact in a variety of ways with DNA and we assert that a similar situation is likely in Zn-finger proteins. Thus, until a 3D model of a Zn-finger protein bound to DNA is available, all of the residues modeled as occurring on the alpha helix away from the beta structure should be considered as primary candidates for variegation when one wishes to alter the DNA-binding properties of a Zn-finger protein. In addition, residues in the beta segment may control interactions with the sugar-phosphate backbone which can effect both specific and non-specific binding.

Parraga et al. (PARR88) have reported a low-resolution structure of a single zinc-finger from NMR data. They confirm the alpha helix proposed by Berg and by Gibson et al., but not the antiparallel beta sheet. The models proposed by Klug and colleagues (FAIR86) have a common feature that is at variance with the models of Berg and of Gibson et al., viz. that the protein chain exits each finger domain at the same end that it entered. The structure published by Parraga et al. does not settle this point, but suggests that the exit strand tends toward the end opposite from the entrance strand, thereby supporting the overall models of Berg and of Gibson et al. Parraga et al. also report that a) a chimeric molecule consisting of zinc-finger domains linked to beta-galactosidase binds sequence-specifically to DNA and b) a protein comprising only two finger motifs can bind sequence-specifically to DNA. They do not suggest that the residues could be mutagenized to achieve novel recognition.

A protein composed of a series of zinc fingers offers the greatest potential of uniquely recognizing a single site in a large genome. A series of zinc fingers is not so well suited to development of a DBP that is sensitive to an effector molecule as is a more compact globular protein such as E. coli CRP. Positive control of genes adjacent to the target DNA subsequence can be achieved as in the case of TF-IIIA.

Overview: Variegation Strategy

Choice of residues in parental potential-DBP to vary:

We choose residues in the initial potential-DBP to vary through consideration of several factors, including: a) the 3D structure of the initial DBP, b) sequences homologous to the initial DBP, c) modeling of the initial DBP and mutants of the initial DBP, d) models of the 3D structure of the target DNA, and e) models of the complex of the initial DBP with DNA. Residues may be varied for several reasons, including: a) to establish novel recognition by changing the residues involved directly in DNA contacts while keeping the protein structure approximately constant, b) to adjust the positions of the residues that contact DNA by altering the protein structure while keeping the DNA-contacting residues constant, c) to produce heterodimeric DBPs by altering residues in the dimerization interface while keeping DNA-contacting residues constant, and d) to produce pseudo-dimeric DBPs (see below) by varying the residues that join segments of dimeric DBPs while keeping the DNA-contacting residues and other residues fixed.

If a dimeric protein comprises two identical polypeptide chains related by a two-fold axis of rotation, we speak of a homodimer with two-fold dyad symmetry. When two very similar polypeptides fold into similar domains and associate, we may observe that there is an approximate two-

fold rotational axis that relates homologous residues, such as the alpha1-beta1 dimer of haemoglobin. We refer to such a pr tein as a heterodimer and to the symmetry axis as a quasi-dyad. When we produce a single-chain DBP by fusing gene fragments that encode two DNA-binding domains joined by a linker amino acid subsequence, we call the molecule a pseudo-dimer and the axis that relates pairs of residues a pseudo-dyad.

Principles that guide choice of residues to vary:

A key concept is that only structured proteins exhibit specific binding, i.e. can bind to a particular chemical entity to the exclusion of most others. In the case of polypeptides, the structure may require stabilization in a complex with DNA. The residues to be varied are chosen to preserve the underlying initial DBP structure or to enhance the likelihood of favorable polypeptide-DNA interactions. The selection process eliminates cells carrying genes with mutations that prevent the DBP from folding. Genes that code for proteins or polypeptides that bind indiscriminately are eliminated since cells carrying such proteins are not viable. Although preservation of the basic underlying initial DBP structure is intended, small changes in the geometry of the structure can be tolerated. For example, the spatial relationship between the alpha 3 helix in one monomer of λ Cro and the alpha 3 helix in the dyad-related monomer (denoted alpha 3') is a candidate for variation. Small changes in the dimerization interface can lead to changes of up to several Å in the relative positions of residues in alpha 3 and alpha 3'.

Burial of hydrophobic surfaces so that bulk water is excluded is one of the strongest forces driving the folding of macromolecules and the binding of proteins to other molecules. Bulk water can be excluded from the region between two molecules or between two portions of a single molecule only if the surfaces are complementary.

The double helix of B-DNA allows most of the hydrophobic surface nucleotides to be buried. The edges of the bases have several hydrogen-bonding groups; the methyl group of thymine is an important hydrophobic group in DNA (HARR88). To achieve tight binding, the shape of the protein must be highly complementary to the DNA, all or almost all hydrogen-bonding groups on both the DNA and the protein must make hydrogen bonds, and charged groups must contact either groups of opposite charge or groups of suitable polarity or polarizability.

There are two complementary interfaces of major interest: a) the DNA-protein interface and b) the interface between protein monomers of dimers or between domains of pseudo-dimers. The DNA-protein interface is more polar than most protein-protein interfaces, but hydrophobic amino acids (e.g. F, L, M, V, I, W, Y) occur in sequence-specific DNA-protein interfaces. The protein-protein interfaces of natural DBPs are typical protein-protein interfaces.

Amino acids are classified as hydrophilic or hydrophobic (ROSE85, EISE86a,b), and although this classification is helpful in analyzing primary protein structures, it ignores that the side groups may contain both hydrophobic and hydrophilic portions, e.g., lysine. Hydrogen bonds and other ionic interactions have strong directional behavior, while hydrophobic interactions are not directional. Thus substitution of one hydrophobic side group for another hydrophobic side group of similar size in an interface is frequently tolerated and causes subtle changes in the interface. For the purposes of the present invention, such hydrophobic-interchange substitutions are made in the protein-protein interface of DBPs so that a) the geometry of the two monomers in the dimer will change, and b) compensating interactions produce exclusively heterodimers.

The process claimed here tests as many surfaces as possible to select one as efficiently as possible that

binds to the target. The selection isolates cells producing those proteins that are more nearly complementary to the target DNA, or proteins in which intermolecular or intramolecular interfaces are more nearly complementary to each other so that the protein can fold into a structure that can bind DNA. The effective diversity of a variegated population is measured by the number of different surfaces, rather than the number of protein sequences. Thus we should maximize the number of surfaces generated in our population, rather than the number of protein sequences. Proteins do not have distinct, countable surfaces; therefore, we define an interaction set as a collection of residues of a protein that can simultaneously touch the target DNA.

If N spatially separated residues of a protein are varied, 20 x N surfaces are generated. Variation of N residues in the same interaction set yields 20^N surfaces. For example, if N = 6, variation of spatially separated residues yields 120 surfaces while variation of interacting residues yields $20^6 = 6.4 \times 10^7$ surfaces. The process of varying residues in an interaction set to maximize the number of surfaces obtained is referred to as Structure-directed Mutagenesis.

If the protein residues to be varied are close enough together in sequence that the variegated DNA (vgDNA) encoding all of them can be made in one piece, then cassette mutagenesis is picked. The present invention is not limited to a particular length of vgDNA that can be synthesized. With current technology, a stretch of 60 amino acids (180 DNA bases) can be spanned.

Mutation of residues further than sixty residues apart can be achieved using other methods, such as single-stranded-oligonucleotide-directed mutagenesis (BOTS85) and two or more mutating primers.

To vary residues separated by more than sixty residues, two cassettes may be mutated serially. From 2-fold to 1000-fold variegation is first introduced into a first cassette. We then introduce 1000-fold to 10^6-fold variegation into a second cassette of the variegated vector population. The composite level of variation preferably does not exceed the prevailing capabilities to a) produce very large numbers of independently transformed cells or b) select small components in a highly varied population. The limits on the level of variegation are discussed below.

Assembly of Relevant Data:

Here we assemble the data about the initial DBP and the target that are useful in deciding which residues to vary in the variegation cycle:

1) 3D structure, or at least a list of residues that contact DNA and that are involved in the dimer contact of the initial DBP,

2) list of sequences homologous to the initial DBP, and

3) model of the target DNA sequence.

These data and an understanding of the function and structure of different amino acids in proteins will be used to answer three questions:

1) which residues of the initial DBP are on the outside and close enough together in space to touch the target DNA simultaneously?

2) which residues of the initial DBP can be varied with high probability of retaining the underlying initial DBP structure?

3) which residues of the initial DBP can affect the dimerization or folding of the initial DBP?

Although an atomic model of the target material is preferred in such examination, it is not necessary.

Graphical and computational tools:

The most appropriate method of picking the residues of the protein chain at which the amino acids should be varied is by viewing with interactive computer graphics a model of the initial DBP complexed with operator DNA. A model based on X-ray data from the DNA-protein complex is preferred, but other models may be used. A stick-figure representation of molecules is preferred. Suitable programs for viewing and manipulating protein and nucleic acid models include: a) PS-FRODO, written by T. A. Jones (JONE85) and distributed by the Biochemistry Department of Rice University, Houston, TX; and b) PROTEUS, developed by Dayringer, Tramantano, and Fletterick (DAYR86). Any hardware that supports either of these programs is appropriate.

Use of Knowledge of Mutations Affecting Protein Stability

In choosing the residues to vary and the substitutions to be made for such residues, one may make use not only of modelling as described above but also of experimental data concerning the effects of mutation in the initial DNA-binding protein. Mutations which will markedly reduce protein stability are to be avoided in most cases.

Missense mutations that decrease DNA-binding protein function non-specifically by affecting protein folding are distinguished from binding-specific mutations primarily on the basis of protein stability (NELS83, PAKU86, VERS86b, HECH84, HECH85a, and HECH85b).

Tables 1, 12, and 13 summarize the results of a number of studies on single missense mutations in the three bacteriophage repression proteins: λ repressor (Table 12) (NELS83, GUAR82, HECH85a, and NELS85), λ Cro (Table 1) (PAKU86, EISE85), and P22 Arc repressor (Table 13) (VERS86a, VERS86b). The majority of the mutant sequences shown in Tables 1, 12, and 13 were obtained in experiments designed to detect loss of function in vivo. The second-site pseudo-reversion mutations (HECH85a), and suppressed nonsense mutations (NELS83), restore function, and some of the site specific changes (EISE85) produce functional proteins.

Roughly 50-70% of the single missense mutations of the DNA-binding proteins selected for loss of function (Tables 1, 12, and 13) produce protein folding defects.

Use of Knowledge of Mutations Affecting the DNA-Protein Interface

Missense mutations in residues thought to be involved in specific interactions with DNA have been reported for several prokaryotic repressor proteins. Table 14 shows an alignment of the H-T-H DNA-binding domains of four prokaryotic repressor proteins (from top to bottom: λ repressor, λ Cro, 434 repressor and trp repressor) and indicates the positions of missense mutations in residues that are solvent-exposed in the free protein but become buried in the protein-DNA complex, and that affect DNA binding.

Randomly obtained missense mutations in solvent-exposed residues of λ repressor, λ Cro, and trp repressor, yield sets of mutants that reduce DNA binding (Table 14). These sets correlate well to the sets of residues that are proposed to interact directly with DNA. Some mutations in λ Cro (EISE85) and all those shown for 434 repressor (WHAR85a) were obtained through site-directed mutagenesis.

Most of the mutations shown in the λ and trp repressor sequences are trans-dominant when the mutant gene is present on an overproducing plasmid (NELS83, KELL85). The exceptions to trans-dominance are the λ repressor SP35 and the trp repressor AT80 mutations. This latter change produces a repressor that has only slightly reduced binding (KELL85). The trans-dominance observed for these mutations is proposed by the authors to result from the wild-type repressor and the mutant repressor forming mixed oligomers which are inactive in binding to operator sites.

Wharton (WHAR85a) has reported that extensive site-directed mutagenesis of 434 repressor positions 28 and 29 produced no functional protein sequences other than the wild-type. Apparently, in the context of 434 repressor structure and operators, only proteins with the wild-type Q28-Q29 sequence bind to the wild-type operators.

Table 14 also shows missense mutations that result in near normal repressor activity. Substitution of 434 repressor Q33 with H, L, V, T, or A produces repressors that function if expressed from overproducing plasmids (WHAR85a); repressor specificity is, however, reduced. Mutations in λ repressor, QY33 (NELS83, HECH83), and in λ Cro, YF26 (EISE85), produce altered proteins which make one less H-bond to the DNA and which bind to the operator DNA with reduced affinity. Thus, loss of a single H-bond is insufficient to completely abolish binding of DNA. Mutations YK26 and HR35 in λ Cro show nearly normal binding (EISE85).

Nelson and Sauer (NELS85) and Hecht et al. (HECH85a, b) have described four replacements in λ repressor (Table 12): EK34, GN48, GS48, and EK83. These derivates have higher affinity for O_R1 than w.t. λ repressor.

Extended amino acid arms at N- and C-terminal locations are important DNA-binding structures in at least four

prokaryotic repressors: λ repr ssor and Cro, and P22 Arc and Mnt.

Sequence-specific and sequence-independent contacts are made by the first 6 amino acid residues (STKKKP) of the λ repressor N-terminal region which form an "arm" that can wrap around the DNA (ELIA85, PABO82a). Missense mutations KE4 and LP12 (Table 12) both greatly reduce repressor activity in vivo (NELS83). Deletion of the first six residues results in a protein which is non-functional in vivo (ELIA85). Deletion of the first three residues results in decrease of affinity for O_R1, loss of protection of back side guanines, altered specificity between O_R1 and O_R3, and decreased binding sensitivity to changes in temperature or salt concentration (ELIA85, PABO82a).

Missense mutations of P22 Arc that produce non-functional proteins with high intracellular specific protein levels (Table 13) are found only in the N-terminal 10 residues of the protein (VERS86b). A single residue change at position 6 (HP6) in P22 Mnt changes operator recognition in the altered protein (YOUD83, VERS86a,b). Knight and Sauer (cited in VERS86a,b) replaced the first 6 residues of Mnt repressor with the first 9 residues of Arc repressor to produce a repressor that binds to the arc operator but not to the mnt operator. Thus P22 Mnt and Arc use a recognition region located in the first 6-10 amino-terminal residues for DNA recognition and binding. The N-terminal DNA-binding of these proteins can not be the recognition helix of a typical H-T-H motif.

In λ Cro, a C-terminal sequence (K62-K63-T64-T65-A66) has been suggested on the basis of model building (TAKE85) and NMR measurements (LEIG87) to form a flexible arm that interacts with minor groove phosphates. Eisenbeis and Caruthers (cited in KNIG88) have found that T64, T65, and A66 have minor effects on protein-operator affinity, while K63 is very important. The C-terminal sequence of P22 Mnt

(K79-K80-T81-T82) is almost identical to that of λ Cro. It has been shown (KNIG88) that deletion of the three residues after K79 has little effect on protein structure or DNA binding. Deletion of K79 and the distal residues, however, reduces operator binding by three orders of magnitude with little apparent change in protein structure.

Use of Knowledge of Mutations Affecting the Protein-Protein Interface

It is also possible to modulate DNA-binding specificity by altering the protein-protein interface. Because the oligomerization equilibrium is coupled to DNA binding, mutations that alter oligomerization affect operator site affinity. Since oligomerization involves the matching of protein surfaces, many interactions are hydrophobic and mutations which specifically destabilize oligomerization are similar to mutations which destabilize global protein structure. Interactions at the site of oligomerization can influence the strength of interactions at the DNA-binding site by subtle alterations in protein structure.

Use of Mutations That Affect Activation

When λ, 434, and P22 repressors bind to their respective O_R2 sites, they activate transcription (POTE80, POTE82, PTAS80). The site on λ repressor which activates RNA polymerase is located on the N-terminal domain of the molecule (BUSH88, HOCH83, SAUE79). Activation requires contact between the N-terminal domain of repressor at O_R2 and RNA polymerase (HOCH83, SAUE79) and this contact stimulates isomerization of the polymerase complex to the open form (McClure and Hawley, cited in GUAR82).

Missense mutations in λ, P22, or 434 repressors that specifically reduce P_{RM} activation while leaving operator binding intact are in the s lvent-exposed protein surface closest to RNA polymerase bound at P_{RM} (GUAR82, PABO79,

BUSH88, WHAR85a). For λ and 434 repressor this surface includes residues in alpha helix 2 and in the turn between alpha helices 2 and 3. In P22 repressor, the surface is formed at the carboxyl terminus of alpha helix 3 (PABO79, TAKE83). In each repressor, the changes that reduce transcriptional activation at P_{RM} involve the substitution of a basic residue for a neutral or acid residue. Further, missense mutations in λ and 434 repressors which increase transcription at P_{RM} involve the substitution of an acidic residue for a neutral or basic residue (GUAR82, BUSH88).

Transcriptional activation at P_{RM} involves the apposition of a negatively charged surface on the N-terminal domain of λ, 434, or P22 repressor to a site on RNA polymerase (BUSH88). Mutations that a) alter the negatively-charged surface of repressor by removing acidic residues or by replacing them with basic residues, or b) that position the negative surface incorrectly with respect to RNA polymerase, decrease transcriptional activation at P_{RM}. Alterations that produce a more negatively charged surface act to increase transcription at P_{RM}.

Pick principal set of residues to vary:

A huge number of variant DNA sequences can be generated by synthesis with mixed reagents at chosen bases. Usually, it is necessary that the number of variants not exceed the number of independently transformed cells generated from the synthetic DNA. It is efficient, however, to make the number of variants as close as practical to this limit. The total number of variants is the product of the number of variants at each varied codon over all the variable codons. Thus, we first consider which residues could be varied with an expectation that alteration could affect DNA binding. We then pick a range of amino acids at each variable residue. The total number of variants is the product of these numbers. If the product is too large or too small, we alter the list of

residues and range of variation at each variable residue until an acceptable number is found.

Considering which residues are on the surface of the initial DBP, we pick residues that are close enough together on the surface of the initial DBP to touch a molecule of the target simultaneously without having any initial DBP main-chain atom come closer than van der Waals distance (viz. 4.0 to 5.0 Å center to center) to any target atom. For the purposes of the present invention, a residue of the initial DBP "touches" the target if:

a) a main-chain atom is within van der Waals distance, viz. 4.0 to 5.0 Å, of any atom of the target molecule,
b) the C_{beta} is within a specific distance of any atom of the target molecule so that a side-group atom could make contact with that atom, or
c) there is evidence that altering the residue alters the DNA-binding of the initial DBP.

The residues in the principal set need not be contiguous in the protein sequence. The exposed surfaces of the residues to be varied need not be connected. We prefer only that the amino acids in the residues to be varied all be capable of touching a single copy of the target DNA sequence simultaneously without atoms overlapping.

In addition to the geometrical criteria, we prefer that there be indications that the initial DBP structure will tolerate substitutions at each residue in the principal set of residues. Indications could come from various sources, including homologous sequences and modeling.

Pick a secondary set of residues to vary:

The secondary set comprises those residues not in the primary set that touch residues in the primary set. These residues might be excluded from the primary set because the

residue is : a) internal, b) highly conserved, or c) on the surface, but the curvature of the initial DBP surface prevents the residue from being in c ntact with the target at the same time as one or more residues in the primary set.

Internal residues are frequently conserved and the amino acid type can not be changed to a significantly different type without risk that the protein structure will be disrupted. Nevertheless, some conservative changes of internal residues, such as I to L or F to Y, are tolerated. Such conservative changes affect the detailed placement and dynamics of adjacent protein residues and such variation may be useful to improve the characteristics of DBP binding.

Surface residues in the secondary set are most often located on the periphery of the principal set. Such peripheral residues can not make direct contact with the target simultaneously with all the other residues of the principal set. It is appropriate to vary the charge of some or all of these residues. For example, the variegated codon containing equimolar A and G at base 1, equimolar C and A at base 2, and A at base 3 yields amino acids T, A, K, and E with equal probability.

Choice of residues to vary simultaneously:

The allowed level of variegation determines how many residues can be varied at once; geometry determines which ones. The user may pick residues to vary in many ways; the following is a preferred manner. The user picks the objective of the variegation, vide supra.

The number of residues picked is coupled to the range through which each can be varied. In the first round progressivity is not an issue; the user may elect to produce a level of variegation such that each molecule of

vgDNA is potentially different through, for example, unlimited variegation of 10 codons (20^{10} approx. = 10^{13} different protein sequences). The levels of efficiency of ligation and transformation reduce the number of DNA sequences actually tested to between 10^7 and 10^9. Multiple performances of the process with very high levels of variegation will not yield repeatable results; the user decides whether this is important.

Pick range of variation:

Each varied residue can have a different scheme of variegation, producing 2 to 20 different possibilities. We require that the process be progressive, i.e. each variegation cycle produces a better starting point for the next variegation cycle than the previous cycle produced.

N.B.: Setting the level of variegation such that the parental pdbp and many sequences related to the parental pdbp sequence are present in detectable amounts insures that the process is progressive. If the level of variegation is so high that the frequency of the parental pdbp sequence can not be detected as a transformant, then each round of mutagenesis is independent of previous rounds and there is no assurance of progressivity. This approach can lead to valuable DNA-binding proteins, but multiple repetitions of the process at this level of variegation will not yield progressive results. Excessive variegation is not preferred in subsequent iterations of this process.

Progressivity is not an all-or-nothing property. So long as most of the information obtained from previous variegation cycles is retained and many different surfaces that are related to the parental DBP surface are produced, the process is progressive. If the level of variegation is

so high that the parental dbp gene may not be detected, the assurance of progressivity diminishes. If the probability of recovering the parental DBP is negligible, then the probability of progressive results is also negligible.

An opposing force in our design considerations is that DBPs are useful in the population only up to the amount that can be detected; any excess above the detectable amount is wasted. Thus we produce as many surfaces related to the parental DBP as possible within the constraint that the parental DBP be present as a marker for the detection level.

Mutagenesis of DNA:

We now decide how to distribute the variegation within the codons for the residues to be varied. These decisions are influenced by the nature of the genetic code. When vgDNA is synthesized, variation at the first base of a codon creates a population coding for amino acids from the same column of the genetic code table (Table 16); variation at the second base of the codon creates a population coding for amino acids from the same row of the genetic code table; variation at the third base of the codon creates a population coding for amino acids from the same box. Work with 3D protein structural models may suggest definite sets of amino acids to substitute at a given residue, but the method of variation may require either more or fewer kinds of amino acids be included. For example, substitution of N or Q at a given residue may be wanted. Combinatorial variation of codons requires that mixing N and Q at one location also include K and H as possibilities at the same residue. The present invention does not rely on accurate predictions of the amino acids to be placed at each residue, rather attention is focused on which residues should be varied.

There are many ways to generate diversity in a

protein (RICH86, CARU85, OLIP86). An extreme case is that one or a few residues of the protein are varied as much as possible (inter alia see CARU85, CARU87, RICH86, WHAR85a). We will call this limit "Focused Mutagenesis". When there is no binding between the parental DBP and the target, we preferably pick a set of five to seven residues on the surface and vary each through all 20 possibilities.

An alternative plan of mutagenesis ("Diffuse Mutagenesis") that may be useful is to vary many more residues through a more limited set of choices (VERS86a,b, INOU86 (Ch.15), PAKU86). This can be accomplished by spiking each of the pure nucleotides activated for DNA synthesis (e.g. nucleotide-phosphoramidites) with one or more of the other activated nucleotides. Contrary to general practice, the present invention sets the level of spiking so that only a small percentage (1% to .00001%, for example) of the final product will contain the parental DNA sequence. This will insure that the majority of molecules carry single, double, triple, and higher mutations and, as required for progressivity, that recovery of the parental sequence will be a possible outcome.

Let N_b be the number of bases to be varied, and let Q be the fraction of all DNA sequences that should have the parental sequence, then M, the fraction of the nucleotide mixture that is the majority component, is

$$M = \exp(\log_e(Q)/N_b) = 10^{(\log_{10}(Q)/N_b)}.$$

If, for example, thirty base pairs on the DNA chain were to be varied and 1% of the product is to have the parental sequence, then each mixed nucleotide substrate should contain 86% of the parental nucleotide and 14% of other nucleotides. Table 17 shows the fraction (fn) of DNA molecules having n non-parental bases when 30 bases are synthesized with reagents that contain fraction M of the majority component. When M=.63096, f24 and higher are less

than 10^{-8}. Note that substantial probability for 8 or more substitutions occurs only if the fraction of parental sequence (f0) drops to around 10^{-3}.

The N_b base pairs of the DNA chain that are synthesized with mixed reagents need not be contiguous. They are picked so that between $N_b/3$ and N_b codons are affected to various degrees. The residues picked for mutation are picked with reference to the 3D structure of the initial DBP, if known. For example, one might pick all or most of the residues in the principal and secondary set. We may impose restrictions on the extent of variation at each of these residues based on homologous sequences or other data. The mixture of non-parental nucleotides need not be random, rather mixtures can be biased to give particular amino acid types specific probabilities of appearance at each codon. For example, one residue may contain a hydrophobic amino acid in all known homologous sequences; in such a case, the first and third base of that codon would be varied, but the second would be set to T. This Diffuse Mutagenesis will reveal the subtle changes possible in the protein backbone associated with conservative interior changes, such as V to I, as well as some not so subtle changes that require concomitant changes at two or more residues of the protein.

Focused Mutagenesis:

If we have no information indicating that a particular amino acid or class of amino acid is appropriate, we approximate substitution of all amino acids with equal probability because representation of one or a few pdbp genes above the detectable level is unproductive. Equal amounts of all four nucleotides at each position in a codon yields the amino acid distribution:

4/64 A	2/64 C	2/64 D	2/64 E	2/64 F	4/64 G
2/64 H	3/64 I	2/64 K	6/64 L	1/64 M	2/64 N
4/64 P	2/64 Q	6/64 R	6/64 S	4/64 T	4/64 V
1/64 W	2/64 Y	3/64 stop			

This distribution has the disadvantage of giving two basic residues for every acidic residue. Such predominance of basic residues is likely to promote sequence-independent DNA binding. In addition, six times as much R, S, and L as W or M occur for the random distribution. Use of equimolar C and G at the third base reduces the over-representation of S, R, and L, but does not cure the maldistribution of acidics and basics.

Consider the distribution of amino acids encoded by one codon in a population of vgDNA. Let Abun(x) be the abundance of DNA sequences coding for amino acid x. For any distribution, there will be a most-favored amino acid (mfaa) with abundance Abun(mfaa) and a least-favored amino acid (lfaa) with abundance Abun(lfaa). We seek the nucleotide distribution that allows all twenty amino acids and that yields the largest ratio Abun(lfaa)/Abun(mfaa) subject to two constraints. First, the abundances of acidic and basic amino acids should be equal. Second, the number of stop codons should be kept as low as possible. Thus only nucleotide distributions that yield

$$\text{Abun(E)+Abun(D)} = \text{Abun(R)+Abun(K)}$$

are considered, and the function maximized is:

$$f(\text{distribution}) = \{(1-\text{Abun(stop)})\ (\text{Abun(lfaa)}/\text{Abun(mfaa)})\}.$$

We limit the third base to equimolar T and G (C and G would be equivalent). All amino acids are possible and the number of accessible stop codons is reduced.

A computer program, "Find Optimum vgCodon." (Table 18), varies the c mposition at bases 1 and 2, in steps of 0.05, and reports the composition that gives the largest value of f(distribution) subject to the constraints:

$$g2 = (g1*a2 - 0.5*a1*a2)/(c1 + 0.5*a1),$$
$$t1 = 1 - a1 - c1 - g1, \text{ and}$$
$$t2 = 1 - a2 - c2 - g2 \quad .$$

The first constraint requires equal amount of acidic and basic amino acids and the second and third conserve matter.

We vary a1, c1, g1, a2, and c2 and then calculate t1, g2, and t2. Initially, variation is in steps of 5%. Once an approximately optimum distribution of nucleotides is determined, the region is further explored with steps of 1%. The optimum distribution is:

Optimum vgCodon

	T	C	A	G
base #1 =	0.26	0.18	0.26	0.30
base #2 =	0.22	0.16	0.40	0.22
base #3 =	0.5	0.0	0.0	0.5

and yields DNA molecules encoding each type of amino acid with the abundances shown in Table 19.

The actual nucleotide distribution obtained in synthetic DNA will differ from the specified nucleotide distribution due to several causes, including: a) differential inherent reactivity of nucleotide substrates, and b) differential deterioration of reagents. It is possible to compensate partially for these effects, but some residual error will occur. We denote the average discrepancy between specified and observed nucleotide fraction as S_{err},

S_{err} = square root (average[$(f_{obs} - f_{spec})/f_{spec}$])

where f_{obs} is the amount of one type of nucleotide found at a base and f_{spec} is the amount of that type of nucleotide that was specified at the same base. The average is over all specified types of nucleotides and over a number (e.g. 10 to 50) of different variegated bases. By hypothesis, the actual nucleotide distribution at a variegated base will be within 5% of the specified distribution. Actual DNA synthesizers and DNA synthetic chemistry may have different error levels. It is the user's responsibility to determine S_{err} for the DNA synthesizer and chemistry employed by the user.

To determine the possible effects of errors in nucleotide composition on the amino acid distribution, we modified the program "Find Optimum vgCodon" in four ways:

1) the fraction of each nucleotide in the first two bases is allowed to vary from its optimum value times (1-S_{err}) to the optimum value times (1 + S_{err}) in seven equal steps (S_{err} is the hypothetical fractional error level), maintaining the sum of nucleotide fractions for one codon position at 1.0,

2) g2 is varied in the same manner as a2, i.e. we dropped the restriction that Abun(D) + Abun(E) = Abun(K) + Abun(R),

3) t3 and g3 are varied from 0.5 times (1 - S_{err}) to 0.5 times (1 + S_{err}) in three equal steps,

4) the smallest ratio Abun(lfaa)/Abun(mfaa) is sought.

In actual experiments, we direct the synthesizer to produce the optimum DNA distribution "Optimum vgCodon" given ab ve. Incomplete control over DNA chemistry may, however, cause us to actually obtain the following distri-

bution that is the worst that can be obtained if all nucleotide fractions are within 5% of the amounts specified in "Optimum vgCodon". A corresponding table can be calculated for any given S_{err} using the program "Find worst vgCodon within S_{err} of given distribution." given in Table 20.

Optimum vgCodon, worst 5% errors

	T	C	A	G
base #1 =	0.251	0.189	0.273	0.287
base #2 =	0.209	0.160	0.400	0.231
base #3 =	0.475	0.0	0.0	0.525

This distribution yields DNA encoding each of the twenty amino acids at the abundances shown in Table 21.

Each codon synthesized with the distribution of bases shown above displays 4 x 4 x 2 = 2^5 = 32 possible DNA sequences, though not in equal abundances. An oligonucleotide containing N such codons would display 2^{5N} possible DNA sequences and would encode 20^N protein sequences. Other variegation schemes produce different numbers of DNA and protein sequences. For example, if two bases in one codon are varied through two possibilities each, then there are 2 x 2 = 4 DNA sequences and 2 x 2 = 4 protein sequences.

If five codons are synthesized with reagents mixed so as to produce the nucleotide distribution "Optimum vgCodon", and if we actually obtained the nucleotide distribution "Optimum vgCodon, worst 5% errors", then DNA sequences encoding the mfaa at all of the five codons are about 277 times as likely as DNA sequences encoding the lfaa at all of the five codons. Further, about 24% of the DNA sequences will have a stop codon in one or more of the five codons.

Consider variegation of a hypothetical sequence, F24-G25-D26-E27-T28, in which each variegated codon is synthesized as an "Optimal vgCodon". The actual abundance of the DNA encoding each type of amino acid is, however, taken from the case of S_{err} = 5% given in Table 21. The abundance of DNA encoding the parental amino acid sequence is:

Amount(parental seq.)

F24	G25	D26	E27	T28
= Abun(F)	* Abun(G)	* Abun(D)	* Abun(E)	* Abun(T)
= .0249	x .0663	x .0545	x .0602	x .0437

$= 2.4 \times 10^{-7}$

Therefore, if the efficiency of the entire process allows us to examine 10^7 different DNA sequences, DNA encoding the parental DBP sequence as well as very many related sequences will be present in sufficient quantity to be detected and we are assured that the process will be progressive.

Setting level of variegation:

We use the following procedure to determine whether a given level of variegation is practical:

1) from: a) the intended nucleotide distribution at each base of a variegated codon, and b) S_{err} (the error level in mixed DNA synthesis), calculate the abundances of DNA sequences coding for each amino acid and stop,

2) calculate the abundance of DNA encoding the parental DBP sequence by multiplying the abundances of the parental amino acid at each variegated residue,

The abundances used in the procedure above are calculated from the worst distribution that is within S_{err} of the specified distribution. A variegation that insures that the parental DBP sequence can be recovered is practical.

Such a level of variegation produces an enormous number of multiple changes related to the parental DBP available for selection of improved successful DBPs. We adjust the subset of residues to be varied and levels of variegation at each residue until the calculated variegation is within bounds.

Reduction of gratuitous restriction sites:

If the method of mutagenesis to be used is replacement of a cassette, we consider whether the variegation generates gratuitous restriction sites. We reduce or eliminate gratuitous restriction sites by appropriate choice of variegation pattern and silent alteration of codons neighboring the sites of variegation.

Focused mutagenesis:

In the preferred embodiment of this process, the number of residues and the range of variation at each residue are chosen to maximize the number of DNA binding surfaces, to minimize gratuitous restriction sites, and to assure the recovery of the initial DBP sequence. For example, in Detailed Example 1, the initial DBP is λ Cro. One primary set of residues includes G15, Q16, K21, Y26, Q27, S28, N31, K32, H35, A36, and R38 of the H-T-H region (Table 14b) and C-terminal residues K56, N61, K62, K63, T64, T65, and A66. A secondary set of residues includes L23, G24, and V25 from the turn portion of the H-T-H region, buried residues T20, A21, A30, I31, A34, and I35 from alpha helices 2 and 3, and dimerization region residues E54, V55, F58, P59, and S60.

The initial set of 5 residues for Focused Mutagenesis contains residues in or near the N-terminal half of alpha helix 3: Y26, Q27, S28, N31, and K32. Varying these 5 residues through all 20 amino acids produces 3.2×10^6 different protein sequences encoded by 32^5 ($=3.3 \times 10^7$)

different DNA sequences. Since all 5 residues are in the same interaction set, this variegation scheme produces the maximum number of different surfaces. Assuming optimized nucleotide distribution described above and S_{err} = 5%, the probability of obtaining the parental sequence is 3.2 x 10^{-7}. This level is within bounds for synthesis, ligation, transformation, and selection capable of examining 10^8 sequences of vgDNA. Codons for the 5 residues picked for Focused Mutagenesis are contained in the 51 bp PpuMII to BglII fragment of the rav^+ gene constructed in Detailed Example 1.

Repetition to obtain desired degree of DNA-binding:

The first variegation step can produce one or more DBPs having DNA-binding properties that are satisfactory to the user. If the best selected DBP is not fully satisfactory, parental DBPs for a second variegation step are picked from DBPs isolated in the first variegation step. The second and subsequent variegation steps may employ either Focused or Diffuse Mutagenesis procedures on residues of the primary or secondary sets. In the preferred embodiment of this process, the user chooses residues and mutagenesis procedures based on the structure of the parental DBP and specific goals. For example, consider three hypothetical cases.

In a first case, a variegation step produces a DBP with greater non-specific DNA binding than is desired. Information from sequence analysis and modeling is used to identify residues involved in sequence independent interactions of the DBP with DNA in the non-specific complex. In the next variegation step, some or all of these residues, together with one or more additional residues from the primary set, are chosen for Focused Mutagenesis and additional residues from the primary or secondary sets are ch sen for Diffuse Mutagenesis.

In a second hypothetical case, a variegation step produces a DBP with strong sequence specific binding to the target and the goal is to optimize binding. In this case, the next variegation step employs Diffuse Mutagenesis of a large number of residues chosen mostly from the secondary set.

In the third hypothetical case, a DBP has been isolated that has insufficient binding properties. A set of residues is chosen to include some primary residues that have not been subjected to variation, one or more primary residues that have been varied previously, and one or more secondary residues. Focused Mutagenesis is performed on this set in the next variegation step.

Overview: DNA Synthesis, Purification, and Cloning

DNA sequence design:

The present invention is not limited to a single method of gene design. The idbp gene need not be synthesized in toto; parts of the gene may be obtained from nature. One may use any genetic engineering method to produce the correct gene fusion, so long as one can easily and accurately direct mutations to specific sites. In all of the methods of mutagenesis considered in the present invention, however, it is necessary that the DNA sequence for the idbp gene be unique compared to other DNA in the operative cloning vector. If the method of mutagenesis is to be replacement of subsequences coding for the potential-DBP with vgDNA, then the subsequences to be mutagenized must be bounded by restriction sites that are unique with respect to the rest of the vector. If single-stranded oligonucleotide-directed mutagenesis is to be used, then the DNA sequence of the subsequence coding for the initial DBP must be unique with respect to the rest of the vector.

The coding portions of genes to be synthesized are

designed at the protein level and then encoded in DNA. The amino acid sequences are chosen to achieve various goals, including: a) expression of initial DBP intracellularly, and b) generation of a population of potential-DBPs from which to select a successful DBP. The ambiguity in the genetic code is exploited to allow optimal placement of restriction sites and to create various distributions of amino acids at variegated codons.

Organization of gene synthesis:

The present invention is not limited as to how a designed DNA sequence is divided for easy synthesis. An established method is to synthesize both strands of the entire gene in overlapping segments of 20 to 50 nucleotides (THER88). An alternative method that is more suitable for synthesis of vgDNA is similar to methods published by others (OLIP86, OLIP87, AUSU87, KARN84). Contrary to most previous workers, we: a) use two synthetic strands, and b) do not cut the extended DNA in the middle. Our goals are: a) to produce longer pieces of dsDNA than can be synthesized as ssDNA on commercial DNA synthesizers, and b) to produce strands complementary to single-stranded vgDNA. By using two synthetic strands, we remove the requirement for a palindromic sequence at the 3' end. Moreover, the overlap should not be palindromic lest single DNA molecules prime themselves.

The present invention is not limited to any particular method of DNA synthesis or construction. Preferably, DNA is synthesized on a Milligen 7500 DNA synthesizer (Milligen, Bedford, MA) by standard procedures. Synthetic DNA is purified by polyacrylamide gel electrophoresis (PAGE) or high-pressure liquid chromatography (HPLC). The present invention is not limited to any particular method of purifying DNA for genetic engineering.

IDBP Gene cloning:

We clone the idbp gene using plasmids that are transformed into competent bacterial cells by standard methods (MANI82) or slightly modified standard methods. DNA fragments derived from nature are operably linked to other fragments of DNA.

Cells transformed with the plasmid bearing the complete idbp gene are tested to verify expression of the initial DBP. Selection for plasmid presence is maintained on all media, while selections for DBP^+ phenotypes are applied only after growth in the presence of inducer appropriate to the promoter. Colonies that display the DBP^+ phenotypes in the presence of inducer and DBP^- phenotypes in the absence of inducer are retained for further genetic and biochemical characterization. The presence of the idbp gene is initially detected by restriction enzyme digestion patterns characteristic of that gene and is confirmed by sequencing.

The dependence of the $IDBP^+$ and $IDBP^-$ phenotypes on the presence of this gene is demonstrated by additional genetic constructions. These are a) excision of the idbp gene by restriction digestion and closure by ligation, and b) ligation of the excised idbp gene into a plasmid recipient carrying different markers and no dbp gene. Plasmids obtained by excising the gene confer the DBP^- phenotypes (e.g. Tc^R, Fus^S, and Gal^S in Detailed Example 1). Plasmids obtained from ligation of idbp to a recipient plasmid confer the DBP^+ phenotypes in the presence of an inducer appropriate to the regulatable promoter (e.g. Tc^S, Fus^R, and Gal^R in Detailed Example 1). Finally, a most important demonstration of the successful construction involves determination of the quantitative dependence of the selected phenotypes on the exogenous inducer concentration.

Overview: DNA-binding Protein Purification and Characterization

Isolation of IDBP:

We purify IDBP and its derivatives by standard methods, such as those described in JOHN80, TAKE86, LEIG87, VERS85b, KADO86.

Quantitation and characterization of protein-DNA binding:

Methods that can be used to quantitate and characterize sequence-specific and sequence-independent binding of a DBP to DNA include: a) filter-binding assays, b) electrophoretic mobility shift analysis, and c) DNase protection experiments. Ionic strength, pH, and temperature are important factors influencing DBP binding to DNA. Standard conditions should correspond closely to the anticipated conditions of use. Thus, if a binding protein is intended for use in bacterial cells in standard culture, a reasonable range of values from which to choose standard conditions would be: pH=7.5 to 8.0, 0.1 to 0.2 M KCl, and 32° to 37°C. Assay buffers preferably include cofactors, stabilizing agents, and counter ions for proper DBP function.

We prepare DNA fragments for analysis of protein-DNA binding by methods that are very similar to those described in MAXA77, KLEN70, RIGB77, and KIMJ87. Filter-binding assays can yield thermodynamic (K_D) and kinetic (k_a and k_d) constants and are performed by methods similar to those described by RIGG70, and KIMJ87. Electrophoretic mobility shift measurements can also yield values of K_D, k_a, and k_d and are performed by methods similar to those of FRIE81. DNase protection assays use the methods of JOHN79, MAXA77, FOXK88. We use chemical methods to characterize binding of proteins to DNA similar to the methods described in BRUN87, BUSH85, and JENJ86.

Table of Examples

--- *** ---

DETAILED EXAMPLE 1

Below is a hypothetical example of a protocol for developing a new DNA-binding protein derived from λ Cro with affinity for a DNA sequence found in human immunodeficiency virus type 1 (HIV-1) using E. coli K-12 as the cell

line or strain. Further optimization, in accordance with the teachings herein, may be necessary to obtain the desired results. Possible modifications in the preferred method are discussed following various steps of the example.

By hypothesis, we set the following technical capabilities:

Yield from DNA synthesis
500 ng/synthesis of ssDNA 100 bases long,
10 ug/synthesis of ssDNA 60 bases long,
1 mg/synthesis of ssDNA 20 bases long.

Maximum oligonucleotide
100 bases

Yield of plasmid DNA
1 mg/l of culture medium

Efficiency of DNA Ligation
0.1 % for blunt-blunt,
4 % for sticky-blunt,
11 % for sticky-sticky.

Yield of transformants
5×10^8 / ug DNA

Error in mixed DNA synthesis (S_{err})
5%

Choice of cell line or strain:

In this example, the following E. coli K-12 recA strains are used: ATCC #35,882 delta4 (Genotype: W3110 trpC, recA, rpsL, sup^o, delta4 (gal-chlD-pgl-att_{lambda}) and ATCC #33,694 HB101 (Genotype: F^-, leuB, proA, recA,

thi, ara, lacY, galK, xyl, mtl, rpsL, supE, hsdS, (r_B^-, m_B^-). E. coli K-12 strains are grown at 37°C in LB broth (MANI82, p440) and on LB agar (addition of 15 g Bacto-agar) for routine purposes. Selections for plasmid uptake and maintenance are performed with addition of ampicillin (Ap) (200 ug/ml), tetracycline (Tc) (12.5 ug/ml) and kanamycin (Km) (50 ug/ml).

Choice of initial DBP:

The initial DBP is λ Cro. Helix-turn-helix proteins are preferred over other known DBPs because more detail is known about the interactions of these proteins with DNA than is known for other classes of natural DBP. λ Cro is preferred over λ repressor because it has lower molecular weight. Cro from 434 is smaller than λ Cro, but more is known about the genetics and 3D structure of λ Cro. An X-ray structure of the λ Cro protein has been published, but no X-ray structure of a DNA-Cro complex has appeared. A mutant of λ Cro, Cro67, confers the positive control phenotype in vitro but not in vivo. The contacts that stabilize the Cro dimer are known, and several mutations in the dimerization function have been identified (PAKU86).

By the methods disclosed herein, DBPs may be developed from Cro which recognize DNA binding sites different from the λ O_R3 or λ operator consensus binding sites, including heterodimeric DBPs which recognize non-symmetric DNA binding sites.

Selections for phenotypes conferred by DBP^+ function:

Media generally are supplemented with IPTG and antibiotic for selection of plasmid maintenance. Cell background is generally strain delta4 (galK,T,E deletion).

a. Galactose resistance (Gal^R). Galactose epimerase deficient ($galE^-$) strains of E. coli (BUTT63) lyse when

treated with galactose. Selective medium is supplemented with 2% galactose, added after autoclaving. Additional galactose, up to 8%, somewhat reduces the background of artifactual galactose-sensitive colonies.

b. Galactose resistance selected immediately after transformation. Inducer IPTG is added to transformed cells, to 5 x 10^{-4} M at the start of the growth period, that allows expression of plasmid antibiotic-resistance. At 60 min after heat shock, cells are further diluted 10-fold into fresh LB broth containing IPTG, antibiotic to select for plasmid uptake (e.g. Ap or Km), and 2% galactose. Cells are grown until lysis is complete or for 3 h, whichever occurs first, then centrifuged at 6,000 rpm for 10 min, resuspended in the initial volume of the post-transformation growth culture, and applied to medium for further selection.

c. Fusaric acid resistance (Tc^S, Fus^R). Successful repression of tet yields resistance to lipophilic chelating agents such as fusaric acid (Fus^R phenotype). Medium described by MALO81 is used for selection of fusaric acid resistance in E. coli; the amount of fusaric acid may be varied. Total cell inoculum is not greater than 5 x 10^6 per plate.

d. Fusaric acid resistance and galactose resistance. Galactose at a final concentration of 2% is added to the medium described by MALO81 after autoclaving. Cells selected directly for galactose resistance in liquid following transformation are applied to this medium.

Selections for phenotypes conferred by DBP^- function:

Cell background is generally strain HB101 ($galK^-$). Media are generally supplemented with IPTG and antibiotic for plasmid maintenance.

a. Tc resistance. Medium, usually LB agar, is supplemented with Tc after autoclaving. Tc stock solution is 12.5 mg/ml in ethanol. It is stored at -20° C, wrapped in aluminum foil. Petri plates containing Tc are also wrapped in foil. Minimum inhibitory concentration is 3.1 ug/ml using a cell inoculum of 5 x 10^7 to 10^8 per plate. More stringent selections employ upto 50 ug/ml Tc. When used for selection of plasmid maintenance, Tc concentration is 12.5 ug/ml.

b. Galactose utilization. Minimal A Medium (MILL72, p432), with galactose as carbon source: after autoclaving add (per liter) 1 ml 1 M M_gSO_4, 0.5 ml of 10 mg/ml thiamine HCl, 10 ml of 20% galactose, and amino acids as required. Cell inoculum per plate is less than 5 x 10^7.

c. Tc resistance and galactose utilization. Medium A with galactose (section b. above) is supplemented with Tce at 3.1 ug/ml.

Selectable systems for DBP isolation:

The tet gene from pBR322 and the E. coli galT,K genes are used in a gal deletion host strain for selection of DBP function. pKK175-6 (BROS84; Pharmacia, Piscataway, NJ), a pBR322 derivative, contains the replication origin, bla (confers Ap^R) for selection of plasmid maintenance, and tet, one of the two selectable genes (Figure 3.) In pKK175-6, tet is promoterless, and all DNA upstream of the pBR322 tet coding region that potentially allow transcription in both directions (BROS82) have been deleted and replaced by the M13 mp8 polylinker. The polylinker and tet are flanked by strong transcription terminators from E. coli rrnB. tet is placed under control of the Tn5 neo promoter, P_{neo}.

Plasmid pAA3H (figure 4) (ATCC #37,308) (AHME84) provides the second set of selectable genes, galT,K. In gal deleted hosts (such as strain ATCC #35,882 carrying the delta4 deletion (E. coli delta4)) plasmid pAA3H confers the Ap^R Tc^S Gal^S phenotype (AHME84) because part of galE is deleted. The galT and galK genes in pAA3H are transcribed from the P_1 "antitet" promoter (BROS82). In E. coli strains carrying galT or galK mutations (e.g. strain HB101), pAA3H confers Gal^+. We place $galT^+$ and $galK^+$ under control of the pBR322 amp gene promoter.

For both tet and gal systems, positive selections are used to select cells that either express or do not express these genes from cultures containing a vast excess of cells of the opposite phenotype.

Placement of test DNA binding sequence:

The test DNA binding sequence for the IDBP, λ O_R3 (KIMJ87), is placed so that the first 5' base is the +1 base of the mRNA transcribed in each of the tet and gal transcription units (Table 100 and Table 101).

Engineering the idbp gene:

A DNA sequence encoding the wild-type Cro protein is designed such that expression is controlled by the lacUV5 promoter. The DNA sequence departs from the wild-type cro gene sequence by the introduction of restriction sites. Thus, the gene is called rav. The transcriptional unit comprising PlacUV5, rav, and trpA terminator is shown in Table 102.

Vector construction:

The construction of an operative cloning vector is summarized in Figure 5. The gal region of pAA3H requires manipulation before insertion into pKK175-6. First the λ-

derived DNA between HpaI and EcoRI is replaced with a ClaI linker (New England BioLabs, #1037). Standard methods are used and the resulting plasmid is named pEP1001 (Figure 6). All plasmids cited in the present application are catalogued in Table 103.

Next, we insert a synthetic fragment, shown in Table 104, comprising the phage fd terminator and two restriction sites (SpeI and SfiI) into the ClaI site of pEP1001; the resulting plasmid is named pEP1002 (Figure 7). Next, we replace the P_1 promoter upstream of gal with P_{amp} from pBR322. As shoen in Table 100, λ O_R3 is positioned downstream of P_{amp} so that it can be used to determine whether binding of Cro can prevent transcription of galT,K. Restriction sites are provided to allow later alteration of the target sequence. The synthetic fragment is cloned into pEP1002 between DraIII and BamHI. The resulting plasmid is named pEP1003 and confers Gal^S on delta4 cells.

The gal genes with the promoter and the fd terminator are moved from pEP1003 into pKK175-6. The 2.69 kb galT,K-bearing HpaI fragment of pEP1003 is ligated to DNA obtained from pKK175-6 by partial DraI digestion. Gal^+ colonies of transformed HB101 cells are picked. The resulting plasmid is named pEP1004 (Figure 9).

The Tn5 neo gene promoter and O_R3 are synthesized (Table 101) and inserted upstream of the tet coding region of pEP1004 between the unique HindIII and SmaI sites. Plasmid DNA from Ap^R Tc^R Gal^S colonies of transformed delta4 cells is analyzed for an insert in the EcoRI-EcoRV fragment of pEP1004. The resulting 7.1 kb plasmid, with two separate selectable gene systems under control of two different promoters and the test DNA binding sequence, is designated pEP1005 (Figure 10).

Cloning the idbp gene:

The BamHI site in the tet gene is removed from the tet gene in pEP1005 by site-directed mutagenesis; the sequence TGG-ATC-CTC that codes for W97-I98-L99 is changed to TGG-ATA-TTG. DNA from pEP1005 is linearized with EcoRV and part (ca. 10%) of the DNA is made single stranded with exonuclease III. The mutagenic oligonuclotide shown in Table 105 is annealed to the DNA that is then completed with Klenow enzyme and ligated. Plasmid DNA from Tc^R, Gal^+ colonies of transformed HB101 is analyzed by standard means; the resulting plasmid is named pEP1006.

Synthetic DNA containing a SpeI overhang, followed by sequences for the lacUV5 promoter, a ribosome binding site, cloning sites for idbp, the trpa terminator (ROSE79), and an SfiI restricted end complementary to the SfiI site in pEP1006 is synthesized as six oligonucleotides as shown in Table 107. We use the methods of THER88 to anneal and ligate these fragments into SpeI, SfiI cut pEP1006. Plasmid DNA from Ap^R, Tc^R, Gal^S colonies of transformed delta4 cells is examined for the SpeI-SfiI insertion by restriction with SpeI, BstEII, BglII, KpnI, and SfiI. The inserted DNA is verified by DNA sequencing, and the 7.22 kb plasmid containing the proper insertion is designated pEP1007, shown in Figure 11.

The idbp gene sequence specifying the Cro^+ protein and designated rav in this Example, is inserted in two cloning steps. The BstEII-BglII segment of rav (Table 109) is inserted first. Oligonucleotides olig#14 and olig#15 are synthesized, annealed, and filled in with Klenow enzyme (Cf. KARN84). The dsDNA is cut wit BstEII and BglII and ligated to BstEII-BglII cut pEP1007. The plasmid containing the appropriate partial rav sequence is designated pEP1008.

The BglII-KpnI fragment of rav is synthesized and inserted in the same manner as the BstEII-BglII fragment. (See Table 110.) This plasmid carrying the complete rav gene is designated pEP1009, shown in Figure 12.

Determine whether IDBP is expressed:

To determine whether cells carrying pEP1009 display the phenotypes expected for rav expression, the delta4 strain bearing pEP1009 is tested on various Ap containing selective media with and without IPTG. Cells are streaked on LB agar media containing: a) Tc; b) fusaric acid; or c) galactose (vide supra). Control strains are the delta4 host with no plasmid, and with pEP1005, pBR322, or pAA3H.

The results below indicate that the rav gene is expressed and the gene product is functional, and that expression is regulated by the lacUV5 promoter.

Growth of derivatives of strain delta4 on selective media (+ Ap)

supplements:	tetracycline		fusaric acid		galactose	
IPTG:	+	-	+	-	+	-
plasmid:						
-	-	-	-	-	-	-
pBR322	+	+	-	-	+	+
pAA3H	-	-	+	+	-	-
pEP1009	-	+	+	-	+	-
pEP1005	+	+	-	-	-	-

λ cI⁻ phage is streaked on each of the above strains, on LB agar with Ap, and with and without IPTG. At sufficiently high intracellular levels of Cro protein, binding of the Cro repressor protein to the λ phage operators O_R and O_L prevents phage growth. Data indicating correct expression and function of the rav gene are:

Growth of λ cI⁻ on delta4 cells

plasmid	phage growth	
	+IPTG	-IPTG
-	+	+
pEP1009	-	+
pEP1005	+	+

These procedures indicate that the chosen IDBP, the product of the rav gene, is expressed and is successfully repressing both the test operators on the plasmid and the wild type operators on the challenge phage.

DBP purification:

Proteins are purified as described by Leighton and Lu (LEIG87).

Quantitation of DBP binding:

We measure DBP binding to the target operator DNA sequence with a filter binding assay, initially using filter binding assay conditions similar to those described for λ Cro (KIMJ87). Data are analyzed by the methods of RIGG70 and KIMJ87.

The target DNA for the assay is the 113 bp ApaI-RsaI fragment from plasmid pEP1009 containing λ O_R3. A control DNA fragment of the same size, used to determine non-specific DNA binding, contains a synthetic ApaI-XbaI DNA fragment specifying the amp promoter and the sequence

5' CTTATACACGAAGCGTGACAA 3' .

This sequence preserves the base content of the O_R3 sequence but lacks several sites of conserved sequence required for λ Cro binding (KIMJ87) and is cloned between

the ApaI-XbaI sites of the pEP1009 backbone to yield pEP1010.

Media Formulations:

Gal^S is demonstrable in LB agar and broth at very low concentrations (0.2% galactose), and is optimal at 2 to 8% galactose. Galactose and Tc selections are performed in LB medium. Fus^S is best achieved in the medium described by Maloy and Nunn (MALO81) for E. coli K-12 strains.

Induction of DBP expression:

The pdbp gene is regulated by the lacUV5 promoter. Optimal induction is achieved by addition of IPTG at 5×10^{-4} M (MAUR80). Experimentation for each successful DBP determines the lowest concentration that is sufficient to maintain repression of the selection system genes.

Optimization of selections

For each selective medium used to detect IDBP function, factors are varied to obtain a maximal number of transformants per plate and with a minimal number of false positive artifactual colonies. Of greatest importance in this optimization is the transcriptional regulation of the initial potential-DBP, such that in further mutagenesis studies, de novo binding at an intermediate affinity is compensated by high level production of DBP.

Regulation of IDBP:

Cells carrying pEP1009 are grown in LB broth with IPTG at 10^{-6}, 5×10^{-6}, 10^{-5}, 5×10^{-5}, 10^{-4} and 5×10^{-4} M. Samples are plated on LB agar and on LB agar containing fusaric acid or galactose as described in above. All media contain 200 ug/ml Ap, and the IPTG concentration of the

broth culture media are maintained in the respective selective agar media.

The IPTG concentration at which 50% of the cells survive is a measure of affinity between IDBP and test operator, such that the lower the concentration, the greater the affinity. A requirement for low IPTG, e.g. 10^{-6} M, for 50% survival due to Rav protein function suggests that use of a high level, e.g. 5×10^{-4} M IPTG, employed in selective media to isolate mutants displaying de novo binding of a DBP to target DNA, will enable isolation of successful DBPs even if the affinity is low.

Concentration of selective agents and cell inoculum size:

Fusaric acid and galactose content of each medium is varied, to allow the largest possible cell sample to be applied per Petri plate. This objective is obtained by applying samples of large numbers of sensitive cells (e.g. 5×10^7, 10^8, 5×10^8) to plates with elevated fusaric acid or galactose. Resistant cells are then used to determine the efficiency of plating. An acceptable efficiency is 80% viability for the resistant control strain bearing pEP1009 in a delta4 background. The total cell inoculum size is increased as is the level of inhibitory compound until viability is reduced to less than 80%.

Choice and cloning of target sequences:

Sequences of the human immunodeficiency virus type 1 (HIV-1) genome were searched for potential target sequences. The known sequences of isolates of HIV-1 were obtained from the GENBANK version 52.0 DNA sequence data base. First we found non-variable regions of HIV-1. We examined the HIV-1 genome from the TATA sequence in the 5'LTR of the HIV-1 genome to the end of the sequence coding for the tat and trs second exons. We intented to locate non-variable regions where a DBP can interfere with the

production of tat and/or trs mRNA because the products of these genes are essential in production of virus (DAYT86, FEIN86).

HIV-1 isolate HXB2 (RATN85) from nucleotide number 1 through 6100 is the reference to which we aligned all other HIV-1 isolates using the Nucleic Acid Database Search program (derived from FASTN (LIPM85)) in the IBI/Pustell Sequence Analysis Programs software package (International Biotechnologies, Inc., New Haven, CT). All stretches of at least 20 bases which have no variation in sequence among all HIV-1 isolates were retained as targets.

From the alignment, segments of the HIV-1 isolate HXB2 sequence that are non-variable among all HIV-1 sequences searched are:

350 - 371, 519 - 545, 623 - 651, 679 - 697,
759 - 781, 783 - 805, 1016 - 1051, 1323 - 1342,
1494 - 1519, 1591 - 1612, 1725 - 1751, 1816 - 1837,
2067 - 2094, 2139 - 2164, 2387 - 2427, 2567 - 2606,
2615 - 2650, 2996 - 3018, 3092 - 3117, 3500 - 3523,
3866 - 3887, 4149 - 4170, 4172 - 4206, 4280 - 4302,
4370 - 4404, 4533 - 4561, 4661 - 4695, 4742 - 4767,
4808 - 4828, 4838 - 4864, 4882 - 4911, 4952 - 4983,
5030 - 5074, 5151 - 5173, 5553 - 5573, 5955 - 5991

In the present Example, these potential regions were searched for subsequences matching the central seven base pairs of the λ operators that have high affinity for λ Cro (viz. O_R3, the symmetric consensus, and the Kim et al. consensus (KIMJ87)). The consensus sequence of Kim et al. has higher affinity for Cro than does O_R3 which is the natural λ operator having highest affinity for Cro. Cro is thought to recognize seventeen base pairs, with side groups on alpha 3 directly contacting the outer four or five bases on each end of the operator. Because the composition and sequence of the inner seven base pairs

affect the position and flexibility of the outer five base pairs to either side, these bases affect the affinity of Cro for the operator.

The sequences sought are shown in Table 111. The letters "A" and "S" stand for antisense and sense. "O_R3A/Symm. Consensus.5" is a composite that has O_R3A at all locations except 5, where it has the symmetric consensus base, C. Similarly, "O_R3A/Symm. Consensus.6" has the symmetric consensus base at location 6 and O_R3A at other locations.

A FORTRAN program searched the non-variable HIV-1 subsequence segments for stretches of seven nucleotides of which at least five are G or C and which are flanked on either side by five bases of non-variable HIV-1 subsequence. The 427 candidate seven-base-pair subsequences obtained using these constraints on CG content were then searched for matches to either the sense or anti-sense strand sequences of the five seven-base-pair subsequences listed above. None of the HIV-1 subsequences is identical to any of the seven-base-pair subsequences. Three HIV-1 subsequences, shown in Table 112, were found that match six of seven bases. Eight subsequences, shown in Table 113, were found that match five out of seven bases and that have five or more GC base pairs. These HIV-1 subsequences are less preferred than the HIV-1 subsequences that match six out of seven bases.

```
                 |
                 11111111
        12345678901234567
5'      aCtTTccGCTggGGaCt   Bases 353-369
        actttccGCTggaaagt   Left symmetrized
        agtcccGCTgggact     Right symmetrized
        tatcAcCGCAAgGgata   OR3
```

(Lower case letters are palindromic in the two halves of the targets and O_R3; highly conserved bases are bold and marked thus a.)

Among the outer five bases of each half operator, bases 1 and 3 are palindromically related to bases 17 and 15 in Target HIV 353-369.

```
                |
TCTCGAcGCAgGACTCG  Bases 681-697
tctcgAcGCAgGcgaga  Left symmetrized
cgagtAcGCAgGactcg  Right symmetrized
tatcAcCGCAAgGgata  O_R3
```

None of bases 1-5 are palindromically related to bases 13-17 in Target HIV 681-697.

```
                |
TTTGAcTAGCGgAGGCT  Bases 760-776
tttgacTAGCGgtcaaa  Left symmetrized
agcctcTAGCGgaggct  Right symmetrized
tatcAcCGCAAgGgata  O_R3
```

None of bases 1-5 are palindromically related to bases 13-17 in Target HIV 760-776.

There is extensive sequence variability among the twelve phage λ operator half-sites. For example:

```
                  |
tAtCaCCGCCGGtGaTa  Consensus
tAtCaCCGCaaGgGaTa  O_R3A
```

The bases in lower case in Consensus and O_R3 sequences shown above are more variable among various lambdoid operators than are bases shown by upper case letters. Studies of mutant operators indicate that A2 and C4 are required for Cro binding. In Target HIV 353-369, bases T3, C6, C7, G8, C9, G14, and A15 match the symmetric consensus sequence, but the highly conserved A2 and C4 are different from lambdoid operators and Cro will not bind to

these subsequences. Mutagenesis of the DNA-contacting residues of alpha 3 is thus the first step in producing a DBP that recognizes the left symmetrized or right symmetrized target sequences.

Target HIV 353-369 is a preferred target because the core (underlined above) is highly similar to the Kim et al. consensus. Target HIV 760-776 is preferred over Target HIV 681-697 because it is highly similar to O_R3.

The method of the present invention does not require any similarity between the target subsequence and the original binding site of the initial DBP. The fortuitous existence of one or more subsequences within the target genes that has similarity to the original binding site of the initial DBP reduces the number of iterative steps needed to obtain a protein having high affinity and specificity for binding to a site in the target gene.

Since the target sequence is from a pathogenic organism, we require that the chosen target subsequence be absent or rare in the genome of the host organism, e.g. the target subsequences chosen from HIV should be absent or rare in the human genome.

Candidate target binding sites are initially screened for their frequency in primate genomes by searching all DNA sequences in the GENBANK Primate directory (2,258,436 nucleotides) using the IBI/Pustell Nucleic Acid Database Search program to locate exact or close matches. A similar search is made of the E. coli sequences in the GENBANK Bacterial directory and in the sequence of the plasmid containing the idbp gene. The sequences of potential sites for which no matches are found are used to make oligonucleotide probes for Southern analysis of human genomic DNA (SOUT75). Sequences which do not specifically bind human DNA are retained as target binding sequences.

The HIV 353-369 left symmetrized and right symmetrized target subsequences are inserted upstream of the selectable genes in the plasmid pEP1009, replacing the test sequences, to produce two operative cloning vectors, pEP1011 and pEP1012, for development of Rav_L and Rav_R DBPs. The promoter-test sequence cassettes upstream of the tet and gal operon genes are excised using StuI-HindIII and ApaI-XbaI restrictions, respectively. Replacement promoter-target sequence cassettes are synthesized and inserted into the vector, replacing O_R3 with the HIV 353-369 left or right symmetrized target sequence in the sequences shown in Table 100 and Table 101.

Choice of residues in Cro to vary:

The choice of the principal and secondary sets of residues depends on the goal of the mutagenesis. In the protocol described here we vary, in separate procedures, the residues: a) involved in DNA recognition by the protein, and b) involved in dimerization of the protein. In this section we identify principal and secondary sets of residues for DNA recognition and dimerization.

Pick principal set for DNA-recognition:

The principal set of residues involved in DNA-recognition is defined as those residues which contact the operator DNA in the sequence-specific DNA-protein complex. Although no crystal structure of a λ Cro-operator DNA complex is available, a crystal structure of a complex between the structural homolog 434 repressor N-terminal domain and a consensus operator has been described (ANDE87). A crystal structure of Cro dimer has been determined (ANDE81) and modeling studies have suggested residues that can make sequence-specific or sequence-independent contacts with DNA in sequence-specific complexes (TAKE83, OHLE83, TAKE85, TAKE86). Isolation and characterization of Cro mutants have identified residues

which contact DNA in protein-operator complexes (PAKU86, HOCH86a,b, EISE85).

Important contacts with DNA are made by protein residues in and around the H-T-H region and in the C-terminal region. Hochschild et al. (HOCH86a,b) have presented direct evidence that Cro alpha helix 3 residues S28, N31, and K32 make sequence-specific contacts with operator bases in the major groove. Mutagenesis experiments (EISE85, PAKU86) and modeling studies (TAKE85) have implicated these residues as well. In addition, these studies suggest that H-T-H region residues Q16, K21, Y26, Q27, H35, A36, R38, and K39 also make contacts with operator DNA. In the C-terminal region, mutagenesis experiments (PAKU86) and chemical modification studies (TAKE86) have identified K56, and K62 as making contacts to DNA. In addition, computer modeling suggests that the 5 to 6 C-terminal amino acids of λ Cro can contact the DNA along the minor groove (TAKE85). From these considerations, we select the following set of residues as a principal set for use in variegation steps intended to modify DNA recognition by Cro or mutant derivative proteins: 16, 21, 26, 27, 31, 32, 35, 36, 38, 39, 56, 62, 63, 64, 65, 66.

Pick secondary set for DNA recognition:

The residues in the secondary set contact or otherwise influence residues in the principal set. A secondary set for DNA recognition includes the buried residues of alpha helix 3: A29, I30, A33, and I34. Interactions between buried residues in alpha helix 2 and buried residues in alpha helix 3 are known to stabilize H-T-H structure and residues in the turn between alpha helix 2 and alpha helix 3 of H-T-H proteins are conserved among these proteins (PTAS86 p102). In λ Cro these positions are T17, T19, A20, L23, G24, and V25. Changes in the dimerization region can influence binding. In λ Cro, residues thought

to be involved in dimer stabilization are E54, V55, and F58 (TAKE85, PABO84). Finally, residues influencing the position of the C-terminal arm of λ Cro are P57, P59, and S60. Thus the secondary set of residues for use in variegation steps intended to modify DNA recognition by λ Cro or Rav proteins is: 17, 19, 20, 23, 24, 29, 30, 33, 34, 54, 55, 57, 58, 59 and 60.

Pick principal set for dimerization:

Different principal and secondary sets of residues must be picked for use in variegation steps intended to alter dimer interactions. In λ Cro, antiparallel interactions between E54, V55, and K56 on each monomer have been proposed to stabilize the dimer (PABO84). In addition, F58 from one monomer has been suggested to contact residues in the hydrophobic core of the second monomer. Inspection of the 3D structure of λ Cro suggests important contacts are made between F58 of one monomer and I40, A33, L23, V25, E54, and A52. In addition, residues L7, I30, and L42 of one monomer could make contact with a large side chain positioned at 58 in the other monomer. Thus, a set of principal residues includes: 7, 23, 25, 30, 33, 40, 42, 52, 54, 55, 56, and 58.

Pick secondary set for dimerization:

The secondary set of residues for variegation steps used to alter dimer interactions includes residues in or near the antiparallel beta sheet that contains the dimer forming residues. Residues in this region are E53, P57, and P59. Residues in alpha helix 1 influencing the orientation of principal set residues are K8, A11, and M12. Residues in the antiparallel beta sheet formed by the beta strands 1, 2, and 3 (see Table 1) in each monomer also influence residues in the principal set. These residues include I5, T6, K39, F41, V50, and Y51. Thus the set of

secondary residues includes: 5, 6, 8, 11, 12, 41, 50, 51, 53, 57, and 59.

Pick the range of variation for alteration of DNA binding:

For the initial variegation step to produce a modified Rav protein with altered DNA specificity a set of 5 residues from the principal set is picked. Focused Mutagenesis is used to vary all five residues through all twenty amino acids. The residues are be picked from the same interaction set so that as many as 3.2×10^7 different DNA binding surfaces will be produced.

A number of studies have shown that the residues in the N-terminal half of the recognition helix of an H-T-H protein strongly influence the sequence specificity and strength of protein binding to DNA (HOCH86a,b, WHAR85, PABO84). For this reason we choose residues Y26, Q27, S28, N31, and K32 from the principal set as residues to vary in the first variegation step. Using the optimized nucleotide distribution for Focused Mutagenesis described above, and assuming that S_{err} = 5% as defined at the start of this Example, the parental sequence is present in the variegated mixture at one part in 3.1×10^5 and the least favored sequence, F at each residue, is present at one part in 10^8. Thus, this level of variegation is well within bounds for a synthesis, ligation, transformation, and selection system capable of examining 5×10^8 DNA sequences.

Pick the range of variation of residues for alteration of dimerization:

As described in the Detailed Description and in this Example, altered λ Cro proteins, Rav_L and Rav_R, that bind specifically and tightly to left and right symmetrized targets derived from HIV 353-369, are first developed through one or more variegation steps. Site-specific changes are then engineered into rav_L to produce dimeriza-

tion defective proteins. Structure-directed Mutagenesis is performed on $\underline{rav}_R$ to produce mutations in Rav_R that can complement dimerization def ctive Rav_L proteins and produce obligate heterodimers that bind to HIV 353-369.

One of the interactions in the dimerization region of λ Cro is the hydrophobic contact between residues V55 of both monomers. The VF55 mutation substitutes a bulky hydrophobic side group in place of the smaller hydrophobic residue; other substitutions at residue 55 can be made and tested for their ability to dimerize. A small hydrophobic or neutral residue present at residue 55 in a protein encoded on expression by a second gene may result in obligate complementation of VF55. In addition, changes in nearby components of the beta strand, E53, E54, K56, and P57 may effect complementation. Thus a set of residues for the initial variegation step to alter the Rav_R dimer recognition is 53, 54, 55, 56, and 57.

Another interaction in the dimerization region of λ Cro is the hydrophobic contact between F58 of one monomer with the hydrophobic core of the other monomer. As mentioned above residues L7, L23, V25, A33, I40, L42, A52, and E54 of one monomer all could make contacts with a large residue at position 58 in the other monomer. The FW58 mutation inserts the largest aromatic amino acid at this position. Compensation for this substitution may require several changes in the hydrophobic core of the complementing monomer. Residues for Focused Mutagenesis in the initial variegation step to alter Rav_R dimer recognition in this case are: 23, 25, 33, 40, and 42.

In each of the two cases described above, the initial variegation step involves Focused Mutagenesis to alter 5 residues through all twenty amino acids. As was shown in Section 6.2.5, this level of variegation is within the limits set by using optimized codon distributions and the

values for S_{err} and transformation yield assumed at the start of this Example.

Mutagenesis of DNA:

Codons encoding λ Cro residues Y26, Q27, S28, N31, and K32 are contained in a 51 bp PpuMI to BglII fragment of the rav gene. To produce the cassette containing the variegated codons we synthesize the 66 nucleotide antisense variegated strand, olig#50, and the primer, olig#52:

```
                d   l   g   v   X   X   X   a   i   X
                22  23  24  25  26  27  28  29  30  31
5' t cct aAG GAC CTA GGG GTG fzk fzk fzk GCG ATT fzk
          ↑ PpuM I↑

 X   a   i   h   a   g   r   k   i
 32  33  34  35  36  37  38  39  40
fzk GCC ATC CAT GCC GGC CGA AAG ATC Tt  3' olig#50
                3'-ccg gct ttc tag aacgccgtg-5' olig#52
                          ↑Bgl II↑
```

The position of the amino acid residue in λ Cro is shown above the codon for the residue. Unaltered residues are indicated by their lower case single letter amino acid codes shown above the position number. Variegated residues are denoted with an upper case, bold X. The restriction sites for PpuMI and BglII are indicated below the sequence. Since restriction enzymes do not cut well at the ends of DNA fragments, 5 extra nucleotides have been added to the 5' end of the cassette. These extra nucleotides are shown in lower case letters and are removed prior to ligating the cassette into the operative vector. The sequence "fzk" denotes the variegated codons and indicates that nucleotide mixtures optimized for codon positions 1, 2, or 3 are to be used. "f" is a mixture of 26% T, 18% C, 26% A, and 30% G, producing four possibilities. "z" is a mixture of 22% T,

16% C, 40% A, and 22% G, producing four possibilities. "k" is an equimolar mixture of T and G, producing two possibilities. Each "fzk" codon produces 4 x 4 x 2 = 2^5 = 32 possible DNA sequences, coding on expression for 20 possible amino acids and stop. The DNA segment above comprises $(2^5)^5 = 2^{25} = 3.2 \times 10^7$ different DNA sequences coding on expression for $20^5 = 3.2 \times 10^6$ different protein sequences.

After synthesis and purification of the variegated DNA, the oligonucleotides #50 and #52 are annealed and the resulting superoverhang is filled in using Klenow fragment as described by Hill (AUSU87, Unit 8.2). The double stranded oligonucleotide is digested with the enzymes PpuMI and BglII and the mutagenic cassette is purified as described by Hill. The mutagenic cassette is cloned into the vectors pEP1011 and pEP1012 which have been digested with PpuMI and BamHI, and the ligation mixtures containing variegated DNA are used to transform competent delta4 cells. The transformed cells are selected for vector uptake and for successful repression at low stringency as described above. Cells containing Rav proteins that bind to the left or right symmetrized targets display the Tc^S, Fus^R and Gal^R phenotypes.

Surviving colonies are screened for correct DBP^+ and DBP^- phenotypes in the presence or absence of IPTG as described above. Relative measures of the strengths of DBP-DNA interactions in vivo are obtained by comparing phenotypes exhibited at reduced levels of IPTG. DBP genes from clones exhibiting the desirable phenotypes are sequenced. Plasmid numbers from pEP1100 to pEP1199 are reserved for plasmids yielding rav_L genes encoding proteins that bind to the Left Symmetrized Targets carried on the plasmids. Similarly, plasmid numbers pEP1200 through pEP1299 plasmids containing rav_R genes encoding proteins that bind to the Right Symmetrized Targets carried on these plasmids.

Based on the determinations above, one or more Rav_L and Rav_R proteins are chosen for further analysis in vitro. Proteins are purified as described above. Purified DBPs are quantitated and characterized by absorption spectroscopy and polyacrylamide gel electrophoresis.

In vitro measurements of protein-DNA binding using purified DBPs are performed as described in the Overview: DNA-Binding, Protein Purification, and Characterization and in this Example. These measurements determine equilibrium binding constants (K_D), and the dissociation (k_d) and association (k_a) rate constants for sequence-specific and sequence-independent DBP-DNA complexes. In addition, DNase protection assays are used to demonstrate specific DBP binding to the Target sequences.

Estimates of relative DBP stability are obtained from measurements of the thermal denaturation properties of the proteins. In vitro measures of protein thermal stability are obtained from determinations of protein circular dichroism and resistance to proteolysis by thermolysin at various temperatures (HECH84) or by differential scanning calorimetry (HECH85b).

One or more iterations of variegation, involving residues thought capable of influencing DNA binding, of the rav_L and rav_R genes produce Rav_L and Rav_R proteins that bind tightly and specifically to the HIV 353-369 left and right symmetrized targets. Additional variegation steps, to optimize protein binding properties can be performed as outlined in the Overview: Variegation Strategy.

By hypothesis, we isolate pEP1127 that contains a pdbp gene that codes on expression for Rav_L-27, shown in Table 114, that binds the left-symmetrized target best among selected Rav_L proteins. Similarly, pEP1238 contains

a pdbp gene that codes on expression for Rav_R-38, shown in Table 115, that binds the right-symmetrized target best among selected Rav_R proteins.

We now use the genes for the Rav_R and Rav_L monomers as starting points for production of obligately heterodimeric proteins Rav_L:Rav_R that recognize the HIV 353-369 target. First we change the target sequences in pEP1238 (containing rav_R-38). We replace both occurrences of the Right Symmetrized Target (in tet and galT,K promoters) with the HIV 353-369 target sequence. Delta4 cells containing plasmids carrying the HIV 353-369 targets display the Ap^R, Tc^R, Fus^S and Gal^S phenotypes. Plasmids carrying HIV 353-369 targets and the rav_R gene are designated by numbers pEP1400 through pEP1499 and corresponding to the number of the donor plasmid of the 1200 series; for example, replacing the target sequences in pEP1238 produces pEP1438.

Engineering dimerization mutants of Rav_L:

To create the site specific VF55 and FW58 mutations in rav_L we synthesize the two mutagenesis primers:

	a	e	e	f	k	p	f		
	52	53	54	55	56	57	58		
5'	GGC	GAA	GAG	TTC	AAG	CCC	TTC	3'	VF55 primer

	v	k	p	w	p	s	n		
	55	56	57	58	59	60	61		
5'	GTA	AAG	CCC	TGG	CCC	AGT	AAC	3'	FW58 primer

Underlining indicates the varied codons and residues. The plasmid pEP1127 (containing rav_L-27) is chosen for mutagenesis. The gene fragment coding on expression for the carboxy-terminal region of the Rav_L protein is transferred into M13mp18 as a BamHI to KpnI fragment. Oligonucleotide-

directed mutagenesis is performed as described by Kunkel (AUSU87, Unit 8.1). The fragment bearing the modified region f Rav_L is removed fr m M13 RF DNA as the BamHI to KpnI fragment and ligated into the correct location in the pEP1100 vector. Mutant-bearing plasmids are used to transform competent cells. Transformed cells are selected for plasmid uptake and screened for DBP^- phenotypes (Tc^R, Fus^S, and Gal^S in E. coli delta4; Gal^+ in E. coli HB101). Plasmids isolated from DBP^- cells are screened by restriction analysis for the presence of the rav_L gene and the site-specific mutation is confirmed by sequencing. The plasmid containing the rav_L-27 gene with the VF55 mutation is designated pEP1301. Plasmid pEP1302 contains the rav_L-27 gene with the FW58 alteration.

For the production of obligate heterodimers as described below, the rav_L^- genes encoding the VF55 or FW58 mutations are excised from pEP1301 or pEP1302 and are transferred into plasmids containing the gene for Km and neomycin resistance (neo, also known as npt II). These constructions are performed in three steps as outlined below. First, the neo gene from Tn5 coding for Km^R and contained on a 1.3 Kbp HindIII to SmaI DNA fragment is ligated into the plasmid pSP64 (Promega, Madison, WI) which has been digested with both HindIII and SmaI. The resulting 4.3 kbp plasmid, pEP1303, confers both Ap and Km resistance on host cells. Next, the bla gene is removed from pEP1303 by digesting the plasmid with AatII and BglI. The 3.5 Kbp fragment resulting from this digest is purified, the 3' overhanging ends are blunted using T4 DNA polymerase (AUSU87, Unit 3.5), and the fragment is recircularized. This plasmid is designated pEP1304 and transforms cells to Km resistance. In the final step, the rav_L^- gene is incorporated in to pEP1304. Plasmid pEP1301 or pEP1302 is digested with SfiI and the resulting 3' overhangs are blunted using T4 DNA polymerase. Next the linearized plasmid is digested with SpeI and the resulting 5' overhangs are blunted using the Klenow enzyme reaction

(KLEN70). The ca. 340 bp blunt-ended DNA fragment containing the entire $\underline{rav_L}^-$ gene is purified and ligated into the PvuII site in pEP1304. Transformed cells are selected for Km^R and screened by restriction digest analysis for the presence of $\underline{rav_L}^-$ genes. The presence of $\underline{rav_L}^-$ genes containing the site-specific VF55 or FW58 mutations is confirmed by sequencing. The plasmid containing the $\underline{rav_L}^-$ gene with the VF55 mutation is designated pEP1305. The plasmid containing the $\underline{rav_L}^-$ gene with the FW58 mutations is designated pEP1306.

In a manner similar to the constructions described above, we ligate the original unmodified $\underline{rav_L}$ gene into pEP1304 to produce plasmid pEP1307.

Engineering heterodimer binding of target DNA:

This round of variegation is performed to produce mutations in Rav_R proteins that complement the dimerization deficient mutations in the Rav_L proteins produced above. To complement the FW58 mutation, the set of five residues L23, V25, A33, I40, and L42 are chosen from the primary set of residues as targets for Focused Mutagenesis.

In an initial series of procedures to test for recognition of HIV 353-369 by the heterodimer Rav_L:Rav_R, we transform cells containing pEP1438 (containing $\underline{rav_R}$-38 and HIV 353-369 targets) with pEP1307 (containing $\underline{rav_L}$). Intracellular expression of $\underline{rav_L}$ and $\underline{rav_R}$ produces a population of dimeric repressors: Rav_L:Rav_L, Rav_L:Rav_R and Rav_R:Rav_R. If the heterodimeric protein is formed and binds to HIV 353-369, cells expressing both $\underline{rav}$ alleles will exhibit the Km^R Ap^R Gal^R Fus^R phenotypes (vide infra). Several pairs of $\underline{rav_L}$ and $\underline{rav_R}$ genes are used in parallel procedures; the best pair is picked for use and further study. Selections for binding the HIV 353-369 target by the heterodimeric protein can be optimized using this system.

Focused Mutagenesis of residues 23, 25, 33, 40, and 42 requires the synthesis and annealing of two overlapping variegated strands because in the rav gene a single cassette spanning these residues extends from the BalI site to the BamHI site and exceeds the assumed synthesis limit of 100 nucleotides. As no variegation affects the overlap, the annealing region is complementary. The antisense strand of the DNA sequence from the BalI site blunt end to the end of the codon for G37 is denoted olig#53.

```
      q   t   k   t   a   k   d   X   g   X   y   q
      16  17  18  19  20  21  22  23  24  25  26  27
5' C CAA ACC AAG ACA GCG AAG GAC fzk GGG fzk TAT CAG
 ↑BalI↑

 s   a   i   n   k   X   i   h   a   g
 28  29  30  31  32  33  34  35  36  37
AGC GCG ATT AAC AAG fzk ATC CAT GCC GGC  3' olig#53
```

f = (26% T, 18% C, 26% A, 30% G)
z = (22% T, 16% C, 40% A, 22% G)
k = equimolar T and G

Olig#53 contains vg codons for residues 23, 25, and 33.

Olig#54 is the sense strand from base 1 in codon 34 to the BamHI site:

```
          i   h   a   g   r   k   X
          34  35  36  37  38  39  40
 3'      TAG GTA CGG CCG GCA TTC jqm

 f   X   t   i   n   a   d   n   k
 41  42  43  44  45  46  47  48  49
AAG jqm TGG TAA TTG CGA CTA CCT AGG cca ca  5' olig#54
                          ↑BamHI↑
```

j = (26% A, 18% G, 16% T, 30% C)
q = (22% A, 16% G, 40% T, 22% C)
m = equimolar A and C

Olig#54 contains variegated codons for residues 40 and 42. Since olig#54 is the sense strand, the variegated nucleotide distributions must complement the distributions for codon positions 1, 2, and 3 used in the antisense strand. These sense codon distributions are designated "j", "q", and "m", and represent the complements to the optimized codon distributions developed for codon positions 1, 2, and 3, respectively, in the antisense strand. The two strands (olig#53 and olig#54) share a 12 nucleotide overlap extending from the first position in the codon for I34 to the end of the codon for G37. The overlap region is 66% G or C.

The two strands shown above are synthesized, purified, annealed, and extended to form dsDNA. Following restriction endonuclease digestion and purification, the mutagenic cassettes are ligated into pEP1438 (containing the asymmetric HIV 353-369 target) in the appropriate locus in the $\underline{rav}_R$ gene. The ligation mixtures are used to transform competent cells that contain pEP1306 (the plasmid with the $\underline{rav}_L$ gene carrying the FW58 site-specific mutation).

Above we picked a set of five residues in λ Cro, E53, E55, V55, K56, and P57, as targets for focused mutagenesis in the first variegation step of the procedure to produce a Rav_R protein that complements the dimerization-deficient VF55 Rav_L mutation. These five residues are contained on a 71 bp BamHI to KpnI fragment of the rav gene (Table 100). To produce a cassette containing the variegated codons we synthesize olig#58:

```
                g   s   v   y   a   X   X   X   X   X   f
                48  49  50  51  52  53  54  55  56  57  58
    5'  ct gat GGA TCC GTC TAC GCG fzk fzk fzk fzk fzk TTC
               ↕BamHI↕

                p   s   n   k   k
                59  60  61  62  63
               CCG AGT AAC AAA AAA

                t   t   a   .
                64  65  66  67
               ACA ACA GCG TAA TAGTAGGTACC ta 3' olig#58
                                  ↕KpnI↕
```

After synthesis and purification of the vgDNA, strands are self-annealed using the 10 nucleotide palindrome at the 3' end of the sequence. The resulting superoverhangs are filled in using the Klenow enzyme reaction as described previously and the double-stranded oligonucleotide is digested with BamHI and KpnI. Purified mutagenic cassettes are ligated into one or more operative vectors (picked from the pEP1200 series) in the appropriate locus in the rav_R gene. The ligation mixtures are used to transform competent cells that contain pEP1305 (the plasmid carrying the rav_L^- gene with the FV55 mutation).

Operative vectors carrying the VF55 or FW58 mutation in rav_L confer Km resistance. Operative vectors carrying mutagenized rav_R genes contain the gene for Ap^R as well as the selective gene systems for the DBP^+ phenotypes. Cells containing complementing mutant proteins are selected by requiring both Ap^R and Km^R and repression of the complete HIV 343-369 target sequence (substituted for the Left and Right Symmetrized Targets in the selection genes). Cells possessing the desired phenotype are Ap^R, Km^R, Fus^R, and Gal^R (in E. coli delta4).

Plasmids from candidate colonies are first isolated genetically by transformation of cells at low plasmid concentration. Cells carrying plasmids coding for Rav_L proteins will be Km^R, while cells carrying plasmids coding for Rav_R proteins will be Ap^R. Plasmids are individually screened to ensure that they confer the DBP^- phenotype and are characterized by restriction digest analysis to confirm the presence of $\underline{rav}_L^-$ or $\underline{rav}_R^-$ genes. Plasmid pairs are co-tested for complementation by restoration of the DBP^+ phenotype when both $\underline{rav}_R$ and $\underline{rav}_L$ are present intracellularly. Successfully complementing plasmids are sequenced through the rav genes to identify the mutations and to suggest potential locations for optional subsequent rounds of variegation.

Plasmids containing genes for altered Rav_R proteins that successfully complement the $\underline{rav}_L$ VF55 mutation are designated by plasmid numbers pEP1500 to pEP1599. Similarly, plasmids containing genes for altered Rav_R proteins that successfully complement the $\underline{rav}_L$ FW58 mutation are designated by plasmid numbers pEP1600 to pEP1699.

Heterodimeric proteins are purified and their DNA-binding and thermal stability properties are characterized as described above. Pairwise variation of the Rav_R and Rav_L monomers can produce dimeric proteins having different dimerization or dimer-DNA interaction energies. In addition, further rounds of variegation of either or both monomers to optimize DNA binding by the heterodimer, dimerization strength or both may be performed.

In this manner a heterodimeric protein that recognizes any predetermined target DNA sequence is constructed. The foregoing is hypothetical. The sequences shown as the result of selection are given by way of example and must not be construed as predictions that proteins of the stated sequence will have specific affinity for any DNA sequence.

--- *** ---

Example 2

Presented below is a hypothetical example of a protocol for developing new DNA-binding polypeptides, derived from the first ten residues of phage P22 Arc and a segment of variegated polypeptide with affinity for DNA subsequences found in HIV-1 using E. coli K12 as the host cell line. Some further optimization, in accordance with the teachings herein, may be necessary to obtain the desired results. Possible modifications in the preferred method are discussed immediately following the hypothetical example.

We set the same hypothetical technical capabilities as used in Detailed Example 1.

Overview:

To obtain significant binding between a genetically encoded polypeptide and a predetermined DNA subsequence, the surfaces must be complementary over a large area, 1000 $Å^2$ to 3000 $Å^2$. For the binding to be sequence-specific, the contacts must be spread over many (12 to 20) bases. An extended polypeptide chain that touches 15 base pairs comprises at least 25 amino acids. Some of these residues will have their side groups directed away from the DNA so that many different amino acids will be allowed at such residues, while other residues will be involved in direct DNA contacts and will be strongly constrained. Unless we have 3D structural data on the binding of an initial polypeptide to a test DNA subsequence, we can not a priori predict which residues will have their side groups directed toward the DNA and which will have their side groups directed outward. We also can not predict which amino acids should be used to specifically bind particular base pairs. Current technology allows production of 10^7 to 10^8

independent transformants per ug of DNA which allows variation of 5 or 6 residues through all twenty amino acids. Alternatively, between 23 and 30 two-way variations of DNA bases can be applied that will affect between 8 and 30 codons.

Sauer and colleagues (VERS87b) have shown that P22 Arc binds to DNA using a motif other than H-T-H. There is as yet no published X-ray structure of Arc, though the protein has been crystallized and diffraction data have been collected (JORD85). A combination of genetics and biochemistry indicates that the first 10 residues of each Arc monomer (M-K-G-M-S-K-M-P-Q-F) bind to palindromically related sets of bases on either side of the center of symmetry of the 21 bp operator shown in Table 200. Furthermore, the first ten residues of each Arc monomer assume an extended conformation (VERS87b). The hydrophobic residues may be involved in contacts to the rest of the protein, but there are several examples from H-T-H DBPs of hydrophobic side groups being in direct contact with bases in the major groove. We do know that these first ten residues of Arc can exist in a conformation that makes sequence-specific favorable contacts with the arc operator.

We pick a target DNA subsequence from the HIV-1 genome such that a portion of the chosen sequence is similar to one half-site of the arc operator. We use part of this chosen sequence for an initial chimeric target. One half of the first target is the DNA subsequence obtained from HIV-1 and the other half of the target is one half-site of the arc operator. For this example, we will use a plasmid bearing wild-type arc operators repressed by the Arc repressor as a control. After demonstrating that Arc repressor can regulate the selectable genes, we replace the wild-type arc operator with the target DNA subsequence. We then replace the arc gene with a variegated pdbp gene and select for cells expressing DBPs that can repress the selectable genes.

Once a protein is obtained that binds t the target that has similarity to one half of the Arc operator, we can change the target so that it has less similarity to one half of the Arc operator and mutagenize those residues that correspond to residues 1-10 of Arc. In vivo selection will isolate a protein that binds to the new target. A few repetitions of this process can producer a polypeptide that binds to any predetermined DNA sequence.

Our potential DNA-binding polypeptide (DBP) will be 36 residues long and will contain the first ten residues of Arc which are thought to bind to part of the half operator. DNA encoding the first ten amino acids of Arc is linked at the 3' terminus of this gene fragment to vgDNA that encodes a further 26 amino acids. Twenty-four of the codons encode two alternative amino acids so that 2^{24} = approx. 1.6×10^7 protein sequences result. The amino acids encoded are chosen to enhance the probability that the resulting polypeptide will adopt an extended structure and that it can make appropriate contacts with DNA. The Chou-Fasman (CHOU78a, CHOU78b) probabilities are used to pick amino acids with high probability of forming beta structures (M, V, I, C, F, Y, Q, W, R, T); the amino acids are grouped into five classes in Table 16. In addition, to discourage sequence-independent DNA binding, some acidic residues should be included. Glutamic acid is a strong alpha helix former, so in early stages we use D exclusively. Further, S and T both can make hydrogen bonds with their hydroxyl groups, but T favors extended structures while S favors helices; hence we use only T in the initial phase. Likewise, N and Q provide similar functionalities on their side groups, but Q favors beta and so is used exclusively in initial phases. Positive charge is provided by K and R, but only R is used in the variegated portion. Alanine favors helices and is excluded. P kinks the chain and is allowed only near the carb xy terminus in initial iterations.

After one selection, we design a different set of binary variegations that includes the selected sequence and perform a second mutagenesis and selection. After two or more rounds of diffuse variegation and selection, we choose a subset of residues and vary them through a larger set of amino acids. We continue until we obtain sufficient affinity and specificity for the target. None of the polypeptides discussed in this example is likely to have a defined 3D structure of its own, because they are all too short. Even if one folded into a definite structure, that structure is unlikely to be related to DNA-binding. A 3D structure, obtained by X-ray diffraction or NMR, of a DNA-polypeptide complex would give us useful indications of which residues to vary. Scattering the variegation along the chain and sampling different charges, sizes, and hydrophobicities produces a series of proteins, isolated by in vivo selection, with progressively higher affinity for the target DNA sequence.

Construction of the test plasmid:

Selection systems are the same as used in Example 1, viz. fusaric acid to select against cells expressing the tet gene and galactose killing by galT,K in a galE deleted host. First, in three genetic engineering steps, we replace: a) the rav gene in pEP1009 with the arc gene, and b) the target DNA sequences (both occurrences) with the arc operator. The resulting plasmid is our wild type control.

To replace rav with arc, the synthetic arc gene, shown in Table 201 and Table 202, is synthesized and ligated into pEP1009 that has been digested with BstEII and KpnI. Cells are transformed and colonies are screened for Tc^R. The plasmid is named pEP2000. Delta4 cells transformed with pEP2000 are Tc^R and Gal^S because pEP2000 lacks the rav gene.

To insert the arc operator into the neo promoter (P_{neo}) for the tet gene in pEP2000, we digest pEP2000 with StuI and HindIII and ligate the purified backbone to annealed synthetic olig#430 and olig#432.

Arc operator and P_{neo} that promotes tet

```
                5'  |CCT|GCG|AAC|CGG|AAT|TGC|CAG|-
Olig #430 =     3'   gga cgc ttg gcc tta acg gtc-
                   | StuI |                | -35  |

|CTG|GGG|CGC|CCT|CTG|GTA|AGG|TTG|-
 gac ccc gcg gga gac cat tcc aac-
                        | -10  |

|GGA|ATG|ATA|GAA|GCA|CTC|TAC|TAT|A        3'=Olig#432
 cct tac tat ctt cgt gag atg ata t tcg a 5'
   |          Arc operator          ||Hind3 |
```

The plasmid is named pEP2001 and confers Fus^R, Gal^S, Ap^R on delta4 cells.

To insert the arc operator into the amp promoter for the galT,K genes in pEP2001, we digest pEP2001 with ApaI and XbaI and ligate the purified backbone to synthetic olig#416 and olig#417 that have been annealed in the standard way.

Arc operator and P_{amp} that promotes galT,K

```
                    5'        |CTT|CTA|AAT|ACA|TTC|AAA|-
          Olig#417  3'  c cgg gaa gat tta tgt aag ttt-
                        | ApaI |              | -35 |

     |TAT|GTA|TCC|GCT|CAT|GAG|ACA|ATA|ACC|-
      ata cat agg cga gta ctc tgt tat tgg-
                          | -10 |

|CTT|ATG|ATA|GAA|GCA|CTC|TAC|TAT| CGT        3'Olig#416
 gaa tac tat ctt cgt gag atg ata  gca gat c  5'
  |       Arc Operator          |    | XbaI |
```

The plasmid is named pEP2002 and confers Gal^R, Fus^R, Ap^R on delta4 cells. This plasmid is our wild type for work with polypeptides that are selected for binding to target DNA subsequences that are related to the arc operator.

Development of polypeptides that bind chimeric target DNA:

We now replace:

a) the two occurrences of the arc operator with the first target sequence that is a hybrid of the arc operator and a subsequence picked from HIV-1, and

b) the arc gene by a variegated pdbp gene.

A hybrid non-palindromic target sequence is used in this example because selection of a polypeptide using a palindromic or nearly palindromic target DNA subsequence is likely to isolate a novel dimeric DBP. The goal of this procedure is to isolate a polypeptide that binds DNA but that does not directly exploit the dyad symmetry of DNA. The binding is most likely in the major groove, but the present invention is not limited to polypeptides that bind in the major groove. The selections are performed using a

non-symmetric target t avoid is lation of novel dimers that support two symmetrically related copies of the original recognition elements.

The non-variable regions of the HIV-1 genome, as listed in Example 1, were searched using a half operator from the arc operator as search sequence.

We sought subsequences in the non-variable sequences of the HIV-1 genome that match either half of the consensus P22 arc operator shown in Table 200. Subsequences that are closer to the start of transcription are preferred as targets because proteins binding to these subsequences will have greater effect on the transcription of the genes. No sequence was found that matched all six unambiguous bases; the subsequences at 1024, 1040, and 2387 (shown in Table 203) each have a single mismatch. Lower case letters in the "arcO =" sequence indicate ambiguity in the P22 arc operator sequence. Lower case, bold, underscored letters in the HIV-1 subsequences indicate mismatch with the consensus arc operator. Two other subsequences, shown in Table 203, have one mismatch at one of the conserved bases and one mismatch with one of the ambiguous bases. The HIV-1 subsequence that starts at base 1024 is chosen as a target sequence. We replace the 3' ten bases of the arc operator with the 3' ten bases of this subsequence to produce the hybrid target sequence:

ATGATAGAAG|C|GCAACCCTC .

We insert this sequence into the promotor that regulates tet in pEP2002 by ligating dsDNA composed of an equimolar mixture of olig#440 and olig#442 into the StuI/HindIII site of pEP2002. Substitution of the arc operator by the arc-HIV-1 hybrid sequence relieves the repression by Arc. The construction is called pEP2003 and confers Tc^R, Ap^R, Gal^S on delta4 cells.

First Target and P_{neo} that promotes tet

```
                5'  |CCT|GCG|AAC|CGG|AAT|TGC|CAG|-
Olig#440 =      3'   gga cgc ttg gcc tta acg gtc-
                   | StuI |              | -35  |

|CTG|GGG|CGC|CCT|CTG|GTA|AGG|TTG|-
 gac ccc gcg gga gac cat tcc aac-
                    | -10  |

    ATA ATA CAG TAg caa ccc tct = HIV 1024-1044
|GGA|ATG|ATA|GAA|GCg|caa|ccc|tct|A       3'=Olig#442
 cct tac tat ctt cgC GTT GGG AGA t tcg a 5'
   |      First Target          ||Hind3 |
```

The second instance of the target is engineered in like manner, using pEP2003 first digested with ApaI and XbaI and then ligated to annealed olig#444 and olig#446. The plasmid is called pEP2004 and confers Gal^+, Tc^R, Ap^R on HB101 cells. The plasmid pEP2004 contains the first target sequence in both selectable genes and is ready for introduction of a variegated pdbp gene.

First Target and P_{amp} that promotes galT,K

```
                5'       |CTT|CTA|AAT|ACA|TTC|AAA|
Olig#444        3'  c cgg gaa gat tta tgt|aag ttt|
                  | ApaI |              | -35  |

|TAT|GTA|TCC|GCT|CAT|GAG|ACA|ATA|ACC|CT-
 ata cat agg cga gta ctc tgt tat tgg ga
                    | -10  |

T|ATG|ATA|GAA|GCg|caa|ccc|tct| CGT       3'Olig#446
a tac tat ctt cgC GTT GGG AGA  gca gat c  5'
|      First Target          |    | XbaI |
```

The variegated DNA for a 36 amino acid polypeptide is shown in Table 204. This DNA encodes the first ten amino acids of P22 Arc followed by 26 amino acids chosen to be likely to form extended structures. In Table 204, we indicate variegation at one base by using a letter, other than A, C, G, or T, to represent a specific mixture of deoxynucleotide substrates. The range of amino acids encoded is written above the codon number:

```
I|M
| 11|
|ATs|
```

indicates that the first base is synthesized with A, the second base with T, and the third base with a mixture of C and G, and that the resulting DNA could encode amino acids I or M. That the parental protein has isoleucine at residue 11 is indicated by writing I first. Residues 22 and 23 are not variegated to provide a homologous overlap region so that olig#420 and olig#421 can be annealed. After olig#420 and olig#421 are annealed and extended with Klenow fragment and all four deoxynucleotide triphosphates, the DNA is digested with both BstEII and Bsu36I and ligated into pEP2004 that has also been digested with BstEII and Bsu36I. The ligated DNA, denoted vg1-pEP2004, is used to transform Delta4 cells. After an appropriate grow out in the presence of IPTG, the cells are selected with fusaric acid and galactose.

By hypothesis, we recover ten colonies that are Gal^R and Fus^R. We sequence the plasmid DNA from each of these colonies. A hypothetical DBP amino acid sequence from one of these colonies is shown in Table 205.

Comparison of the amino-acid sequences of different isolates may provide useful information on which residues

play crucial roles in DNA binding. Should a residue contain the same amino acid in most or all isolates, we might infer that the selected amino acids is preferred for binding to the target sequence. Because we do not know that all of the isolates bind in the same manner, this inference must be considered as tentative. Residues closer to the unvaried section that have repetitive isolates containing the same amino acid are more informative than residues farther away.

In a second round of Diffuse Mutagenesis, we vary the codons shown in Table 206. Residues 1 through 10 are not varied because these provide the best match for the first ten bases of the target. Residues 19, 20, and 21 are not varied so that the synthetic oligonucleotides can be annealed. The two-way variations at residues 11 through 18 and 23 through 36 all allow the selected amino acid to be present, but also allow an as-yet-untested amino acid to appear. It is desirable to introduce as much variegation as the genetic engineering and selection methods can tolerate without risk that the parental DBP sequence will fall below detectable level. Having picked three residues for the homologous overlap, we have only 23 amino acids to vary. Thus residue 22 is varied through four possibilities instead of only two. Residue 22 was chosen for four-way variegation because it is next to the unvaried residues. We use pEP2004 as the backbone, and ligate DNA prepared with Klenow fragment from oligonucleotides #423 and #424 (Table 206) to the BstEII and Bsu36I sites. The resulting population of plasmids containing the variegated DNA is denoted vg2-pEP2004.

Table 207 shows the amino acid sequence obtained from a hypothetical isolate bearing a DBP gene specifying a polypeptide with improved affinity for the target. Changes in amino acid sequence are observed at ten positions. Comparisons of the sequences from several such isolates as well as those obtained in the first round of mutagenesis

can be used to locate residues providing significant DNA-binding energy.

Having established some affinity for the target, we now seek to optimize binding via a more focused mutagenesis procedure. Table 208 shows a third variegation in which twelve residues in the variable region are varied through four amino acids in such a way that the previously selected amino acids may occur. Again, pEP2004 is used as backbone and synthetic DNA having cohesive ends is prepared from olig#325 and olig#327. The plasmid is denoted vg3-pEP2004. In subsequence variegation, we would vary other residues through four amino acids at one time. By hypothesis, we select the polypeptide shown in Table 209 that has high specific affinity for the first target; now we can:

a) replace both occurrences of the first target by a second target, i.e. the intact HIV-1 subsequence (1024-1044), and

b) use the selected polypeptide as the parental DBP to generate a variegated population of polypeptides from which we select one or more that bind to the second target.

Because the second target differs from the first in the region thought to be bound by residues 1 through 10 of the parental DBP, we concentrate our variegation within these residue for the first several rounds of variegation and selection.

We replace the target DNA sequence in the neo promoter for tet in pEP2002 with ds DNA comprising synthetic olig#450 and olig#452 at the StuI/HindIII site. The new plasmid is named pEP2010 and confers Tc^R on delta4 cells.

Second Target and P_{neo} that promotes tet

```
                5'  |CCT|GCG|AAC|CGG|AAT|TGC|CAG|-
Olig#450 =      3'   gga cgc ttg gcc tta acg gtc-
                   | StuI |                | -35 |

|CTG|GGG|CGC|CCT|CTG|GTA|AGG|TTG|GG-
 gac ccc gcg gga gac cat tcc aac cc-
                        | -10 |

 ATA ATA CAG TAg caa ccc tct = HIV 1024-1044
A|ATa|ATA|cAg|tag|caa|ccc|tct|A       3'Olig#452
t taT tat Gtc ATC GTT GGG AGA t tcg a 5'
 |     Second Target        ||Hind3 |
```

We replace the target in the amp promoter for galT,K of pEP2010 with synthetic olig#454 and olig#456 between ApaI and XbaI sites. The new plasmid is named pEP2011 and confers Gal^+ on HB101. pEP2011 contains the second target in both selectable genes and is ready for introduction of a variegated pdbp gene and selection of cells expressing polypeptides that can selectively bind the target DNA subsequence.

Second Target and P_{amp} that promotes galT,K

```
                5'       |CTT|CTA|AAT|ACA|TTC|AAA|
Olig#454        3'  c cgg gaa gat tta tgt|aag ttt|
                   | ApaI |              | -35 |

|TAT|GTA|TCC|GCT|CAT|GAG|ACA|ATA|ACC|CT
 ata cat agg cga gta ctc tgt tat tgg ga
                        | -10 |
```

```
  ATA ATA CAG TAg caa ccc tct = HIV 1024-1044
T|ATa|ATA|cAg|tag|caa|ccc|tct| CGT        3'Olig#456
a taT tat GtC ATc GTT GGG AGA  gca gat c  5'
|     Second Target          |   | XbaI |
```

Variegation of the first eleven residues of the potential DNA-binding polypeptide is illustrated in Table 210. Double-stranded DNA having appropriate cohesive ends is prepared from olig#460 and olig#461, Klenow fragment, BstEII, and Bsu36I. This DNA is ligated into similarly digested backbone DNA from pEP2011; the resulting plasmid is denoted vg1-pEP2011. Delta4 cells are transformed and selected with fusaric acid and galactose. Table 211 shows the sequence of a 37 amino-acid polypeptide isolated from cells exhibiting the DBP^+ phenotypes by the above hypothetical selection. The sequence shown in Table 211 is hypothetical and is given by way of example. This example must not be construed as a prediction that this sequence has specific affinity for the target or any other DNA sequence. Further variegation (vg2, vg3,...) of this peptide and selection for binding to Target#2 will be needed to obtain a peptide of high specificity and affinity for Target#2.

We anticipate that Successful DBP production will take more than three or four cycles of variegation and selection, perhaps 10 or 15. We anticipate that initial phases will require careful adjustment of the selective agents and IPTG because the level of repression afforded by the best polypeptide may be quite low. As stated, we expect that biophysical methods, such as X-ray diffraction or NMR, applied to complexes of DNA and polypeptide will yield important indications of how to hasten the forced evolution.

The length of the polypeptide in the example may not be optimal; longer or shorter polypeptides may be needed. It may be necessary to bias the amin acid composition more

toward basic amino acids in initial phases to obtain some non-specific DNA binding. Inclusion of numerous aromatic amino acids (W,F,Y,H) may be helpful or necessary.

Other strategies to obtain polypeptides that bind sequence-specifically are illustrated in examples 3, 4, and 5.

Example 3

We present a second example of the application of our selection method applied to the generation of asymmetric DBPs. A possible problem with making and using DNA-binding polypeptides, is that the polypeptides may be degraded in the cell before they can bind to DNA. That polypeptides can bind to DNA is evident from the information on sequence-specific binding of oligopeptides such as Hoechst 33258. Polypeptides composed of the 20 common natural amino acids contain all the needed groups to bind DNA sequence-specifically. These are obtained by an efficient method to sort out the sequences that bind to the chosen target from the ones that do not. To overcome the tendency of the cells to degrade polypeptides, we will attach a domain of protein to the variegated polypeptide as a custodian. The first example of a custodial domain presented is residues 20-83 of barley chymotrypsin inhibitor.

The strategy is to fuse a polypeptide sequence to a stable protein, assuming that the polypeptide will fold up on the stable domain and be relatively more protected from proteases than the free polypeptide would be. If the domain is stable enough, then the polypeptide tail will form a make-shift structure on the surface of the stable domain, but when the DNA is present, the polypeptide tail will quickly (a few milliseconds) abandon its former protector and bind the DNA. The barley chymotrypsin

inhibitor (BCI-2) is chosen because it is a very stable domain that does not depend on disulfide bonds for stability. We could attach the variegated tail at either end of BCI-2. A preferred order of amino acid residues in the chimeric polypeptide is: a) methionine to initiate translation, b) BCI-2 residues 20-83, c) a two residue linker, d) the first ten residues of Arc, and e) twenty-four residues that are varied over two amino acids at each residue. The linker consists of G-K. Glycine is chosen to impart flexibility. Lysine is included to provide the potentially important free amino group formerly available at the amino terminus of the Arc protein. The first target is the same as the first target of Example 2.

Table 300 shows the sequence of a gene encoding the required sequence. The ambiguity of the genetic code has been resolved to create restriction sites for enzymes that do not cut pEP1009 outside the rav gene. This gene could be synthesized in several ways, including the method illustrated in Table 301 involving ligation of oligonucleotides 470-479. Plasmid pEP3000 is derived from pEP2004 by replacement of the arc gene with the sequence shown in Table 300 by any appropriate method.

Table 302 illustrates variegated olig#480 and olig#481 that are annealed and introduced into the CI2-arc(1-10) gene between PpuMI and KpnI to produce the plasmid population vg1-pEP3000. Cells transformed with vg1-pEP3000 are selected with fusaric acid and galactose in the presence of IPTG. Further variegation (vg2, vg3, ...) will be required to obtain a polypeptide sequence having acceptably high specificity and affinity for Target#1.

Example 4

We present a second strategy involving a polypeptide chain attached to a custodial domain. In this strategy,

the custodial domain contains a DNA-recognizing element that will be exploited to obtain quicker convergence of the forced evolution.

The three alpha helices of Cro fold on each other. It has not been observed that these helices fold by themselves, but no efforts in this direction have been reported. We will attach a variegated segment of 24 residues to residue 35 of Cro (H35 is the last residue of alpha 3). The target will be picked to contain a good approximation to the half O_R3 site at one end but no constraint is placed on the bases corresponding to the dyad-related other half of O_R3. A sequence that departs widely from the O_R3 sequence is actually preferred, because this discourages selection of a novel dimeric molecule. We assume that alpha-3 forms and binds to the same four or five bases that it binds in O_R3 and that a polypeptide segment attached to the carboxy terminus of alpha-3 can continue along the major groove. We attach 24 amino acids of polypeptide immediately after the last residue of alpha-3, wherein the polypeptide is chosen: a) to have more positive charge than negative charge, b) to have beta chain predominate, c) to have some aromatic groups, and d) to have some H-bonding groups, produces a population that is then cloned and host cells are selected for expression of a polypeptide that binds preferentially to the target sequence.

We first construct a hybrid target sequence (Target #3) containing one O_R3 half-site fused to a portion of the final target. This hybrid target DNA subsequence is inserted into the selectable genes in the same manner as the arc operator was inserted in Example 2. We then follow the same procedure to vary the 24 residues; first we vary twenty-four residues, using two possible amino acids at each residue. We carry out two or more cycles of such diffuse variegation. Then we vary 12 residues, using 4 possible amino acids at each residue. We do two or more

iterations of this process so that all residues are varied at least once.

We have now generated one or more DBPs that bind well to one half of the final target sequence. Next we generate binding to the other half of the final target. First we replace both instances of Target #3 with the final target sequence, target #4. We then vary the alpha helix 3 and the surface of the hypothesized domain formed by helices 1-3 to optimize binding to final target sequence.

A search of the non-variable regions of the HIV-1 genome reveals that bases 624-640 (aATCtCTAGCAGTGGCG) contain a good match to one half of O_R3, as shown in Table 400. As first target of this example, we choose

TATCCCTAGCAGTGGCG,

denoted Target#3, that has one half of O_R3 and nine bases from HIV-1. Once a sequence is obtained that binds Target#3, we replace Target#3 by Target#4 = HIV 624-640 and variegate the recognition helices taken from Cro.

To engineer Target#3 into P_{neo} that regulates tet, plasmid pEP2002 is digested with StuI and HindIII and the purified backbone is ligated to an annealed, equimolar mixture of olig#490 and olig#492. Delta4 cells are transformed and selected with Tc; replacement of the arc operator relieves the repression by Arc. Plasmid DNA from Tc^R colonies is sequenced to confirm the construction; the construction is called pEP4000.

Target #3 and P_{neo} that promotes tet

```
                5'  |CCT|GCG|AAC|CGG|AAT|TGC|CAG|-
Olig#490 =      3'   gga cgc ttg gcc tta acg gtc-
                  | StuI |              | -35  |
```

```
|CTG|GGG|CGC|CCT|CTG|GTA|AGG|TTG|GG-
 gac ccc gcg gga gac cat tcc aac cc-
                    | -10  |
```

```
 aAT CtC TAG CAG TGG CG = HIV 624-640
A|TAT|CCC|TAG|CAG|TGG|CGA        3'Olig#492
t ata ggg atc gtc acc gct tcg a 5'
 |       Target #3     ||Hind3 |
```

We engineer the second instance of the target, in like manner, into P_{amp} for galT,K, using ApaI and XbaI to digest pEP4000 and olig#494 and olig#496. HB101 cells (galK$^-$) are transformed and are selected for ability to grow on galactose as sole carbon source. Plasmid DNA from Gal$^+$ colonies is sequenced in the region of the insert to confirm the construction. The plasmid is called pEP4001.

Target #3 and P_{amp} that promotes galT,K

```
                5'        |CTT|CTA|AAT|ACA|TTC|AAA|
     Olig#494   3'  c cgg gaa gat tta tgt|aag ttt|
                   | ApaI |              | -35  |
```

```
|TAT|GTA|TCC|GCT|CAT|GAG|ACA|ATA|ACC|
 ata cat agg cga gta ctc tgt tat tgg
                    | -10  |
```

```
|CTT|TAT|CCC|TAG|CAG|TGG|CG CGT        3'Olig#496
 gaa ata ggg atc gtc acc gc gca gat c  5'
   |       Target #3     |     | XbaI |
```

A gene fragment encoding the first two helices of Cro is shown in Table 401. Olig#483 and olig#484 are synthesized and extended in the standard manner and the DNA is digested with BstEII and KpnI. This DNA is ligated to

backbone from pEP4001 that has been digested with BstEII and KpnI; the resulting plasmid, denoted pEP4002, contains the Target#3 subsequence in both selectable genes and is ready for introduction of a variegated pdbp gene between BglII and KpnI. Table 402 shows a piece of vgDNA prepared to be inserted into the BglII-KpnI sites of pEP4002. Table 403 shows the result of a selection of delta4 cells, transformed with vg1-pEP4002, with fusaric acid and galactose in the presence of IPTG. Additional cycles of variegation of residues 36-61 are carried out in such a way that the amino acid selected at the previous cycle is included. After several cycles in which 22-24 residues are varied through two possible amino acids, we choose 10-13 amino acids and vary them through four possibilities.

Once reasonably strong binding to Target#3 is obtained, we replace Target#3 with Target#4 and vary the residues in helix 3 (residues 26-35) and, to a lesser extent, helix 2 (residues 16-23).

Example 5

We disclose here a method of engineering a polypeptide extension onto the amino terminus of P22 Arc, a natural DBP, so that the novel DBP develops asymmetric DNA-binding specificity for a subsequence found in the HIV-1 genome. Others have observed that loss of arms from natural DBPs may cause loss of binding specificity and affinity (PABO82a and ELIA85), but none, to our knowledge, have suggested adding arms to natural DBPs in order to enhance or alter specificity or affinity. The new construction is denoted a "polypeptide extension DBP"; the gene is denoted ped and the proteins are denoted Ped. Wild-type Arc forms dimers and binds to a partially palindromic operator. We will generate a sequence of DBPs descendent from Arc. Early members of this family will form dimers, but will have sufficient binding area such that asymmetric targets will

be bound. In final stages of the development, proteins that do not dimerize will be engineered.

Table 200 shows the symmetric consensus of left and right halves of the P22 arc operator, arcO. Table 500a shows a schematic representation of the model for binding of Arc to arcO that is supported by genetic and biochemical data (VERS87b). Arc is thought to bind B-DNA in such a way that residues 1-10 are extended and the amino terminus of each monomer contacts the outer bases of the 21 bp operator (RT Sauer, public talk at MIT, 15 September 1987).

Arc is preferred because: a) one end of the polypeptide chain is thought to contact the DNA at the exterior edge of the operator, and b) Arc is quite small so that genetic engineering is facilitated. P22 Mnt is also a good candidate for this strategy because it is thought that the amino terminal six residues contact the mnt operator, mntO, in substantially the same manner as Arc contacts arcO. Mnt has significant (40%) sequence similarity to Arc (VERS87a). Mnt forms tetramers in solution and it is thought that the tetramers bind DNA while other forms do not. When the mnt gene is progressively deleted from the 3' end to encode truncated proteins, it is observed that proteins lacking K79 and subsequent residues have lowered affinity for mntO and that proteins lacking Y78 and subsequent residues can not form tetramers and do not bind DNA sequence-specifically (KNIG88). Some truncated Mnt proteins of 77 or fewer residues form dimers, but these dimers do not present the DNA-recognizing elements in such a way that DNA can be bound. Arc is preferred over Mnt because Arc is smaller and because Arc acts as a dimer.

Other natural DBPs that have DNA-recognizing segments thought to interact with DNA in an extended conformation (referred to as arms or tails) and thought to contact the central part of the operator, such as λ Cro or λ cI repressor, are less useful. For these proteins to be

lengthened enough to c ntact DNA outside the original operator, several residues would be needed to span the space between the central bases contacted by the existing terminal residues and the exterior edge of the operator.

Table 500a illustrates interaction of Arc dimers with arcO; the two "C"s of Arc represent the place, near residue F10, at which the polypeptide chain ceases to make direct contact with the DNA and folds back on itself to form a globular domain, as shown in Table 500b and Table 500c. Which of these alternative possibilities actually occurs has not been reported. Our strategy is compatible, with some alterations, with either structure. In Table 500b, each set of residues 1-10 makes contact with a domain composed of residues 11-57 of the same polypeptide chain; the dimer contacts are near the carboxy terminus. Table 500c shows an alternative interaction in which residues 1-10 of one polypeptide chains interact with residues 11-57 of the other polypeptide chain; the dimer contacts occur shortly after residue 10. The similarity of sequences of Arc and Mnt, the demonstration of function of DNA-recognizing segments transferred from Arc to Mnt (RT Sauer, public talk at MIT, 15 September 1987 and Knight and Sauer cited in VERS86b), and the behavior of Mnt on truncation suggest that Table 500b is the correct general structure for Arc, but the structure diagrammed in Table 500c is also possible.

Table 501 shows the four sites at which one of the consensus arc half operators comes within one base of matching ten bases (six unambiguous and four having two-fold ambiguity) in the non-variable segments of HIV-1 DNA sequence, as listed in Example 1. The symbol "@" marks base pairs that vary among different strains of HIV-1. Because we intend to extend Arc from its amino terminus, we seek subsequences of HIV-1 that: a) match one of the arc half operators, and b) have non-variable sequences located so that an amino-terminal extension of the Arc protein will

interact with non-variable DNA. Th subsequences 1024-1033 and 4676-4685 meet this requirement while the subsequences at 1040-1049 and 2387-2396 do not. In the case of 1040-1049, the amino-terminal extension would proceed in the 3' direction of the strand shown and would reach variable DNA after two base pairs. For 2387-2396, variable sequence is reached at once. The subsequence 1024-1033 is preferred over the subsequence 4676-4685 because it is much closer to the beginning of transcription of HIV so that binding of a protein at this site will have a much greater effect on transcription. In the remainder of this example, positions within the target DNA sequence will be given the number of the corresponding base in HIV-1. Base A_{1034} of HIV-1 is aligned with the central base of arcO.

HIV 1024-1044 has only three bases in each half that are palindromically related to bases in the other half by rotation about base pair 1034: A_{1024}/T_{1044}, A_{1026}/T_{1042}, and G_{1032}/C_{1036}. The latter two base pairs correspond to positions in arcO that are not palindromically related. Five of the six palindromically related bases of arcO correspond to non-palindromically related bases in HIV 1024-1044. Thus no dimeric protein derived from Arc is likely to bind HIV 1016-1046 if symmetric changes are made only in the residues 1-10 (or in any other set of residues originally found in Arc). Our strategy is to add, in stages, eleven variegated residues at the amino terminus and to select for specific binding to a progression of targets, the final target of the progression being bases 1016-1037 of HIV-1. Because the region of protein-DNA interaction is increased beyond that inferred for wild-type Arc-arcO complexes, unfavorable contacts in bases aligned with the right half of arcO can be compensated by favorable contacts of the polypeptide extension with bases 1016-1023. The penultimate selection isolates a dimeric protein that binds to the HIV-1 target 1016-1037; the ultimate selection isolates a protein that does not dimerize and binds to the same target.

Table 502 shows a progression of target sequences that leads from wild-type arc0 to HIV 1016-1037. It is emphasized that finding a subsequence of HIV-1 that has high similarity to one half of arc0 is not necessary; rather, use of this similarity reduces the number of steps needed to change a sequence that is highly similar to arc0 into one that is highly similar or identical to an HIV-1 subsequence. Reducing the number of steps is useful, because, for each change in target, we must: a) construct plasmids bearing selectable genes that include the target sequence in the promoter region, b) construct a variegated population of ped genes, and c) select cells transformed with plasmids carrying the variegated population of ped genes for DBP^+ phenotype.

In sections (a), (c), (e), and (g) of Table 502, bases in the targets are in upper case if they match HIV 1016-1046 and are underscored if they match the wild-type arc0 sequence.

We construct a series of plasmids, each plasmid containing one of the target sequences in the promoter region of each of the selectable genes. For each target, we variegate the ped gene and select cells for phenotypes dependent on functional DBPs. For each target, several rounds of variegation and selection may be required. We anticipate that a plurality of proteins will be obtained from independent isolates by selection for binding to one target. We pick the protein that shows the strongest in vitro binding to short DNA segments containing the target as the parental Ped to the next round of variegation and selection. Genetic methods, such as generation of point mutations in the ped gene or in the target and selection for function or non-function of Ped can be used to determine associations between particular bases and particular residues (VERS86b).

Once a Ped with specific binding for the target is obtained, it may be useful to determine a 3D structure of the Ped-DNA complex by X-ray diffraction or other suitable means. Such a structure would provide great help in choosing residues to vary to improve binding to a given target or to an altered target.

We initiate development of a polypeptide extension DBP having affinity for HIV 1016-1037 by generating a variegated population of Peds and selecting for binding to the first target. Table 502a shows the first target which we designed to have identity to arc0 in the left half, but to have a mismatch (arc0 vs. target) at A_{1038} (which is C in the corresponding position in the right half of arc0 and is palindromically related to a G in the left half); the rationale is as follows. Vershon et al. (VERS87b) report that chemical modification with dimethyl sulfate of the wild-type CG at this location interferes mildly with binding of Arc and that this location is strongly protected from modification by dimethylsulfate if Arc is bound to the operator. Thus we expect a mismatch between wild-type arc0 and the first target at A_{1038} to make wild-type Arc bind poorly. Binding can be restored, however, by favorable contacts to bases 1021-1023 by the amino-terminal extension.

An alternative first target would have C_{1038}, as does arc0 at the corresponding location, and A_{1041}, unlike arc0 or HIV-1. Vershon et al. (VERS87b) report that methylation of the corresponding CG base pair strongly interferes with binding of Arc. Thus, changing the base that corresponds to HIV 1041 should have a strong effect on binding of Arc to the alternative target.

In the first variegation step, we extend Arc by five variegated residues at the amino terminal. Since five residues can contact no more than three bases in a sequence-specific manner, we limit the extent of the target

to those bases that correspond to HIV 1021-1044. Inclusion of bases corresponding to HIV 1016-1020 at this initial stage might position the target too far downstream from the promoters of the selectable genes to allow strong repression of these promoters. Once a Ped displaying binding to bases corresponding to 1021-1044 has been isolated, we can introduce a greater length of the HIV-1 sequence into the left side of the target without concern that the Ped will bind too far downstream from the promoter of the selectable genes to block transcription. Furthermore, once binding by the amino terminal extension has been established, we can, in a stepwise manner, remove the right half of arcO from the target, thereby forcing more asymmetric binding to the left half of arcO and the bases upstream of 1024.

The first target is engineered into both selectable genes as in Example 2. We use olig#501 and olig#502, shown in Table 503, to introduce the first target downstream of P_{neo} that promotes tet, replacing arcO in pEP2002; the resulting plasmid is called pEP5000. From pEP5000, we use olig#503 and olig#504 to construct pEP5010 in which the first target replaces arcO downstream of P_{amp} that promotes galT,K.

Table 502b shows schematically how the amino terminal residues align to the first target; the five residue extension is unlikely to contact more than 3 base pairs upstream from base 1024. The alteration in the right half operator prevents tight binding unless the additional residues make favorable interactions upstream of 1024. Care is taken in designing the two instances of the target that the downstream boundaries are different, AAG in P_{neo} and CGT in P_{amp}. Thus, for the novel DBP to bind specifically to both instances of the target, it must recognize the common sequence upstream of base 1024.

An initial variegated ped is constructed using olig#605, as shown in Table 504, and comprises: a)

methionine codon to initiate translation, b) five variegated codons that each allow all twenty possible amino acids, and c) the Arc sequence from 101 to 157. (Because we are constructing a polypeptide extension at the amino terminus, we have added 100 to the residue numbers within Arc so that Arc residue 1 is designated 101.) This variegated segment of DNA comprises $(2^5)^5 = 2^{25} = 3.2 \times 10^7$ different DNA sequences and encodes $20^5 = 3.2 \times 10^6$ different protein sequences; with the given technical capabilities, we can detect each of the possible protein sequences. The 3' terminal 20 bases of olig#605 are palindromically related so that each synthetic oligonucleotide primes itself for extension with Klenow enzyme. The DNA is then digested with Bsu36I and BstEII and is ligated to the backbone of appropriately digested pEP5010 which bears the first target in each selectable gene. Transformed delta4 cells are selected for Fus^R Gal^R at low, medium, and high concentrations of IPTG, the inducer of the lacUV5 promoter that regulates ped. Because the first target is quite similar to arcO, we anticipate that a functional Ped will be isolated with low-level induction of the ped gene with IPTG.

More than one round of variegation and selection may be required to obtain a Ped with sufficient affinity and specificity for the first target. Function of a Ped is judged in comparison to the protection afforded by wild-type Arc in cells bearing pEP2002. Specifically, strength of Ped binding is measured by the IPTG concentration at which 50% of cells survive selection with a constant concentration of galactose or fusaric acid, chosen as a standard for this purpose. A Ped is deemed acceptable if it can protect cells against the standard concentrations of galactose and fusaric acid, administered in separate tests, with an IPTG concentration of 5×10^{-4} M. Preferably, a Ped can protect cells against the standard concentrations of galactose and fusaric acid, tested separately, with no more than ten times the concentration of IPTG

needed by pEP2002-bearing cells. Variegation of residues 101, 102, and others may be needed. We anticipate that a plurality of independent functional Peds will be isolated; we discriminate among these by measuring in vitro binding to DNA oligonucleotides that contain the target sequence. The amino-acid sequences of different isolates are compared; residues that always contain only one or a few kinds of amino acids are likely to be involved in sequence-specific DNA binding. Table 505 shows a hypothetical isolate, Ped-6, that binds the first target.

Table 502c shows the changes between the first target and the second target. Three changes are made left of center to make the target more like HIV 1016-1042. Only the change G_{1030}->C affects a base that is palindromically related in arcO. One change is made right of center that makes the target more like HIV 1016-1042, less like arcO, and less palindromically symmetric. Furthermore, the target is shortened on the right by two bases so that selection isolates proteins that bind asymmetrically to the left side of the target. Starting with pEP2002, we introduce, in two genetic engineering steps that use olig#541, olig#542, olig#543 and olig#544 (Table 506), the second target (in place of arcO) into the promoter region of each selectable gene; the resulting plasmid is denoted pEP5020.

Table 507 shows a variegated sequence that is ligated into pEP5020 between BstEII and Bsu36I. Variegated codons are shown in the same way as in Table 204.

Table 502d illustrates that residues 100-110 of Ped-6 contact the bases of the second target that differ from the first target. Accordingly, residues 1 and 96-99 of Ped are not variegated in the DNA shown in Table 507; rather, residues 100-110 are each varied through four possibilities, always including the amino acid previously present at that residue. This generates $4^{11} = 2^{22}$ = approx. 4 x

10^6 different DNA and protein sequences. Selection of transformed delta4 cells for Fus^R Gal^R and screening by in vitro DNA binding yields, by hypothesis, a plasmid coding on expression for the protein Ped-6-2, illustrated in Table 508.

An alternative to the variegation shown in Table 507 is one in which we vary residues 101-105, 108, and 110 through eight possibilities each, yielding 2.0×10^6 DNA and protein sequences. These residues, except M101, are indicated to be in contact with the operator. M101 has been altered by the attachment of the polypeptide extension and thus should be altered. After variegation of the listed residues and selection, further variegation should include some variegation of residues 96-103 because changes in the listed residues may change the context within which residues 96-103 contact the DNA.

More than one round of variegation and selection may be required to obtain a Ped having sufficient affinity and specificity for the second target.

Table 502e shows the changes from the second target to the third, which comprise: a) inclusion of bases 1018-1020, b) one change to the left of the 21 bp arcO region, c) two changes at the center of the arcO region, d) two changes left of center, and e) removal of bases 1041 and 1042. All of these changes make the third target less symmetric and more like HIV 1016-1040. The third target is introduced into each of the selectable genes in the same manner as the second target. The resulting plasmid, obtained in two genetic engineering steps, is denoted pEP5030. Table 502f shows that residues 96-110 are all potential sites to alter the specificity and affinity of DBPs derived from Ped-6-2. Thus, in Table 510, we illustrate a segment of variegated DNA that comprises $2^{20} = 10^6$ DNA sequences and encoding on expression 10^6 protein sequences having ten residues varied through two possibilities and five residues

through four possibilities. The DNA is then digested with BstEII and Bsu36I and ligated into pEP5030. Transformed delta4 cells are selected for Fus^R Gal^R. By hypothesis, we isolate a plasmid, denoted pEP5031, that codes on expression for the protein Ped-6-2-5 shown in Table 509.

Table 502g shows the changes between the third and fourth targets. The changes are: a) inclusion of bases 1016-1017, b) two changes right of center, and c) removal of bases 1038-1040. The initial variegation to be selected using the fourth target consists of an extension of six residues at the amino terminus of Ped-6-2-5, shown in Table 511. In iterative steps of forced evolution of proteins, one should not produce a number of different DNA sequences greater than the number of independent transformants that one can obtain (about 10^8 with current technology). Because there are no residues corresponding to 90-95 in the parental DBP (Ped-6-2-5), the first variegation and selection with the fourth target is a non-iterative step and it is permissible to produce 10^{10} DNA sequences and 6.4×10^7 protein sequences. In subsequent iterative rounds of variegation, the number of variants is, preferably, limited to a fraction, e.g. 10%, of the number of independent transformants that can be generated and subjected to selection. A protein, illustrated in Table 512 and denoted Ped-6-2-5-2, is isolated, by hypothesis, through selection of a variegated population of transformed cells for Fus^R Gal^R.

Ped-6-2-5-2 binds specifically to HIV 1016-1037 as a dimer. HIV 1016-1037 has no palindromic symmetry. Binding to an asymmetric DNA sequence by a dimeric protein is possible because the Ped-6-2-5-2 dimer has more recognition elements than wild-type P22 Arc dimer and so can bind even though nearly half of the right half of arcO has been removed from the target. Ped-6-2-5-2 is useful as is; nevertheless, obtaining a monomeric protein may have advantages, including: a) higher affinity for the target

because suboptimal interactions are eliminated, and b) lower molecular weight. Obtaining a functional monomeric Ped is easiest if Arc dimers interact in the manner shown in Table 500b. We use the following steps to isolate a protein that binds specifically to HIV 1016-1037 as a monomer.

Ped-6-2-5-2 is the parental DBP from which we derive the monomeric DBP. The route taken from a palindromically symmetric arcO sequence to an asymmetric HIV sequence was designed to select for binding to the left half of the original arc operator.

Proteins that do not dimerize, but that bind specifically to the fourth target can be generated in several ways. Because the 3D structure of Arc is still unknown, we can not use Structure-Directed Mutagenesis to pick residues to vary to eliminate dimerization. One way to obtain monomeric proteins is to use diffuse mutagenesis to vary all residues from 111 to 157 and select for proteins that can bind the target sequence. Another strategy is to synthesize the ped gene in such a way that numerous stop codons are introduced. This causes a population of progressively truncated proteins to be expressed. Table 513 shows a segment of variegated DNA that spans the BglII to KpnI sites of the arc gene used throughout this example. This segment is synthesized with suitable spacer sequences on the 5' end. The extra "t" at the 3' end allows two such chains to prime each other for extension with Klenow enzyme. The ratios of bases in the variegated positions are picked so that each varied codon encodes about 35% of polypeptides to terminate at that position. Since we intend to determine how much the protein can be shortened and remain functional, we begin by replacing codon 153 with stop. Since 15 residues are varied, only about 0.3 % of chains will continue to stop codon 153 without one or more stop codons. All the intermediate length chains will be present in the selection in detectable amount. delta4

cells transformed with pEP5030 containing this vgDNA are selected for Fus^R Gal^R. Because each variegated codon causes translation termination in about 35% of the genes in the variegated population, shorter coding regions are more abundant than longer ones. Thus, the shortest gene that encodes a functional repressor will be the most abundant gene selected. Plasmid DNA from a number of independent selected colonies is sequenced. The dimerization properties of several functional DBPs are tested in vitro and the sequence of the shortest monomeric protein is retained for use and further study.

In this manner, we generate a protein that binds monomerically to a DNA sequence that has no palindromic symmetry.

Example 6

We illustrate here the fusion of two known DNA-binding domains to form a novel DNA-binding protein that recognizes an asymmetric target sequence. The progression of targets is the same as shown in Table 502 (Example 5). The amino-acid sequence of the initial DBP is illustrated in Table 600 and comprises the third zinc-finger domain from the product of the Drosophila kr gene (ROSE86), a short linker, and P22 Arc. The linker consists of three residues that are picked to allow: a) some flexibility between the two domains, and b) introduction of a KpnI site. The polypeptide linker should not allow excessive flexibility because this would reduce the specificity of the DBP.

The primary set of residues to vary to alter the DNA-binding are marked with asterisks. Those in the zinc finger were picked by reference to the model of Gibson et al. (GIBS88); all residues having outward-directed side groups (except those directed upward from the beta strands) were picked. Residues 101-110 (1-10 of Arc) were also

picked to be in the primary set. Other residues within the Arc sequence may be varied. For each target in the progression, we initially choose for variegation residues in the primary set that are most likely to abut that part of the target most recently changed. For example, for the first target, we begin by varying residues 21, 24, 25, 28, and 29, each through all twenty amino acids. After one or more rounds of variegation and selection, other residues in the primary and secondary set are varied.

Other zinc-finger domains, such as those tabulated by Gibson et al. (GIBS88), are potential binding domains. Other proteins with known DNA binding, such as 434 Cro, may be used in place of Arc. Multiple zinc fingers could be added, stepwise, to obtain higher levels of specificity and affinity.

TABLES

Table 1 MISSENSE MUTATIONS IN λ CRO

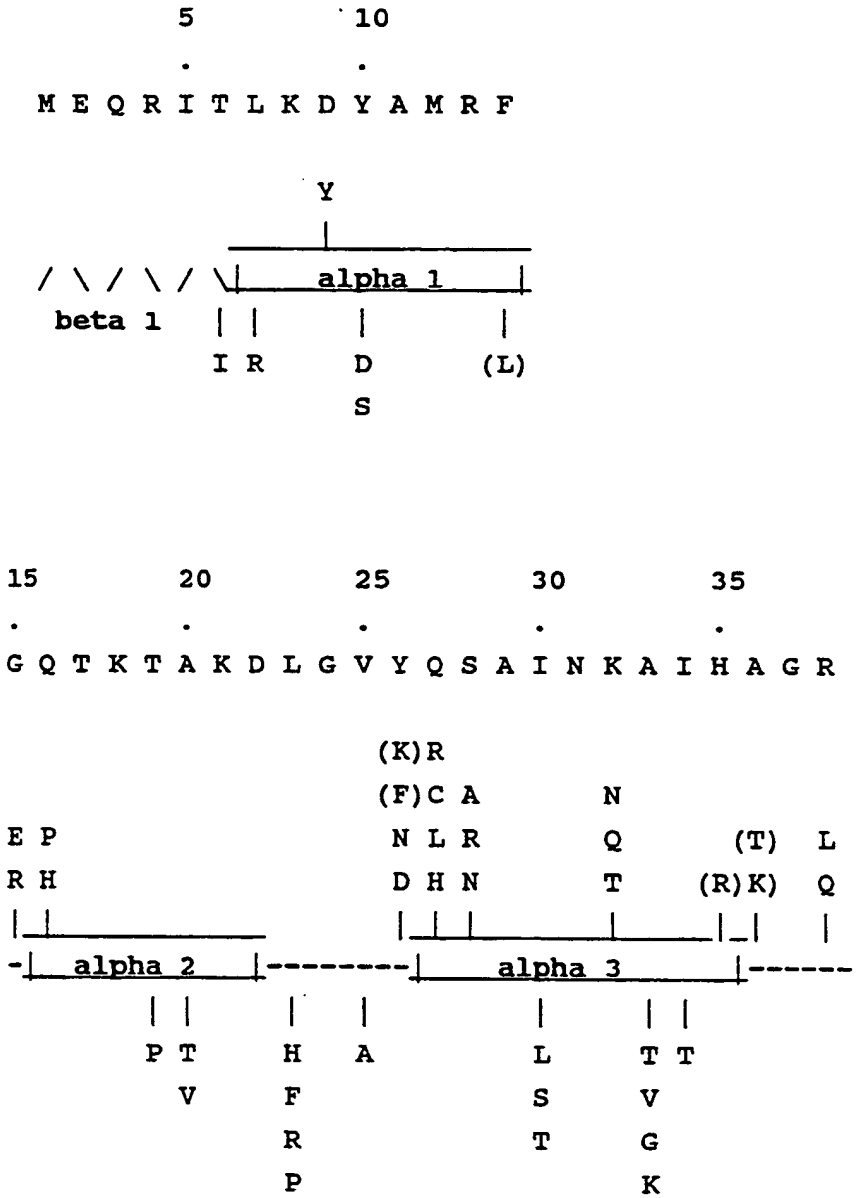
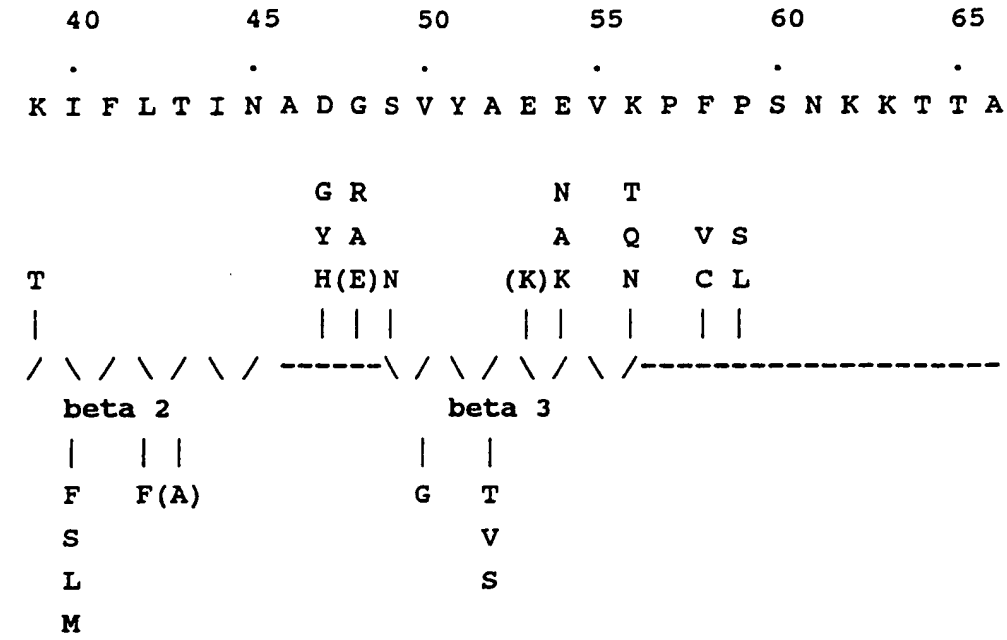


Table 1, continued



Notes:

Substitutions occuring at solvent exposed positions in the unbound repressor dimer are shown above the wild type sequence.

Substitutions occuring at internal positions are shown below the wild type sequence.

Subsitutions that produce repressor dimers with normal or nearly normal DNA binding affinities are shown in parentheses.

Table 2: Examples of selections for plasmid uptake and maintenance in E. coli

gene	(alternate designation)	function
Amp^R	(Ap^R, bla)	beta-lactamase
Kan^R	(Km^R, neo)	aminoglycoside P-transferase
Tet^R	(Tc^R, tet)	membrane pump
Cam^R	(Cm^R, cat)	acetyltransferase
colicin immunity		binds to colicin in vivo
$TrpA^+$		complementation of trpA

Table 3: Examples of selections for plasmid uptake and maintenance in S. cerevisiae

gene	function
$Ura3^+$	complements ura3 auxotroph
$Trp1^+$	complements trp1 auxotroph
$Leu2^+$	complements leu2 auxotroph
$His3^+$	complements his3 auxotroph
Neo^R	resistance to G418

Table 4: Agents for Selection
of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Galactose-1-phosphate uridylyltransferase and galactokinase				
galT$^+$ & *galK*$^+$	galactose	*galE*$^-$, *galT*$^-$, *galK*$^-$	galactose as sole C source	*galE*$^+$ & (*galT*$^-$ or *galK*$^-$)
Tetracycline resistance (*E. coli* K-12 strains are Tet^S)				
tetA$^+$	fusaric acid	Tet^S	Tc	Tet^S
beta galactosidase				
lacZ$^+$	phenylgalactoside	*lacZ*$^-$	lactose as sole C source	*lacZ*$^-$

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Phe tRNA synthetase				
pheS$^+$	fluorophenylalanine	*pheS12*$^+$	growth at high temperature	*pheS*-amber, sup-ts
Transport of arginine, lysine, and ornithine				
argP$^+$	canavanine	*argP*$^-$ Arg prototroph	requirement for arginine and lysine at low conc. in medium	*argP*$^-$ & Arg auxotroph Lys auxotroph

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
thymidylate synthetase				
thyA$^+$	trimethoprim + thymidylate	*thyA*$^-$	thymidylate omitted from defined medium	*thyA*$^-$
cAMP Receptor Protein (note 3)				
crp$^+$	fosfomycin	*crp*$^-$	lactose or other regulated sugar as sole C source	*crp*$^-$
Orotidine-5'-phosphate decarboxylase				
pyrF$^+$	5-fluoroorotate	*pyrF*$^-$	Thymine & cytosine requirement on	*pyrF*$^-$

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
mannosephosphotransferase enzyme II				
ptsM$^+$	deoxyglucose	*ptsM*$^-$	Mannose as sole C source	*ptsM*$^-$
Fusion protein				
secA$^+$ & *malE* signal-*lacZ* fusion	lactose as sole C (note 2)	*secA*$^-$ & *lacZ*$^-$	phenylgalactoside	*secA*$^-$ & *lacZ*$^-$

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Outer membrane protein (note 4)				
ompA$^+$	colicin E1	*ompA*$^-$	HfrH(*thr*$^+$,*leu*$^+$, *str*^{S}) conjugation	*thr*$^-$,*leu*$^-$, *ompA*,*str*^{R}
	colicin E2			
	colicin E3			
	phage TuII			
	phage K3			
	phage 4-59			

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Vitamin B12 transport				
btuB$^+$	phage BF23	*btuB*$^-$ B12 prototroph	requirement for B12 in defined medium	*btuB*$^-$ & B12 auxotroph

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Maltose transport				
lamB$^+$	Phage λ	*lamB*$^-$	growth on maltose as sole C source	*lamB*$^-$
Ferrichrome receptor				
tonA$^+$	Phage phi80	*tonA*$^-$	Requirement for Fe hydroxamate as sole Fe source	*tonA*$^-$
Colicin I receptor				
cir$^+$	colicin I	*cir*$^-$	screen for colicin I resistance	*cir*$^-$

Table 4 (continued): Agents for Selection of DBP Binding in *E. coli* and Relevant Genotypes

Plasmid Genotype	Forward Selection Agent	Forward Selection Host Genotype	Reverse Selection Agent	Reverse Selection Host Genotype
Nucleoside uptake, colicin K receptor, phage T6 receptor (note 5)				
tsx$^+$	colicin K Phage T6	*tsx*$^-$	requirement for nucleosides in defined medium	*tsx*$^-$ thymine auxotroph purine auxotroph
Aromatic amino acid transport				
aroP$^+$	thienylalanine or fluorophenylalanine	*aroP*$^-$	requirement for tryptophan in defined medium	Trp auxotroph *aroP*$^-$
Cysteine synthetase				
cysK$^+$	selenate or azaserine in medium	*cysK*$^-$	growth on medium lacking cysteine	*cysK*$^-$